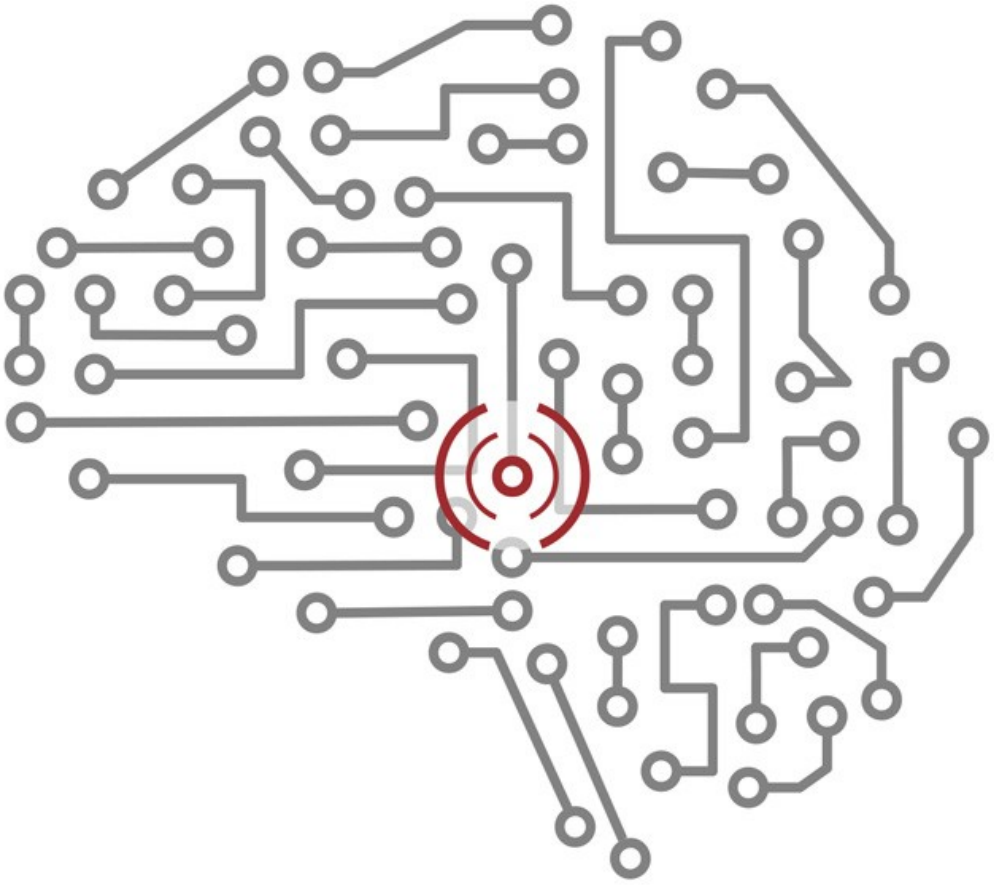
UNIVERZITA KOMENSKÉHO V BRATISLAVE

Peter Šantavý



UMELÁ INTEL!GENCIA

DOBRÝ SLUHA A ZLÝ PÁN?

Ľudia, stroje, spoločnosť...
Etika, hodnoty, budúcnosť...

UNIVERZITA KOMENSKÉHO V BRATISLAVE

RÍMSKOKATOLÍCKA CYRILOMETODSKÁ

BOHOSLOVECKÁ FAKULTA

Peter Šantavý

Umelá inteligencia – dobrý sluha a zlý pán?

Ľudia, stroje, spoločnosť...

Etika, hodnoty, budúcnosť...

Bratislava, 2023

Vydala Rímskokatolícka cyrilometodská bohoslovecká fakulta
Univerzity Komenského v Bratislave v roku 2023.

Prvé vydanie

Rozsah: 309 strán; 18 AH

ISBN 978-80-88696-91-9 (pdf)
EAN 9788088696919

Autor

ThLic. Ing. Peter Šantavý, PhD.

Katolícky kňaz, správca linuxových systémov a počítačových sietí, expert v oblasti kybernetickej bezpečnosti, nechtiac technologický manažér a príležitostný učiteľ IKT (informačných a komunikačných technológií).

Má viac ako 25 rokov skúseností s IKT a je jedným z odborníkov Katolíckej cirkvi v oblasti kybernetickej bezpečnosti, informačnej spoločnosti i komplexných počítačových systémov a sieťových technológií.

Fenoménu umelej inteligencie sa na pôde Univerzity Komenského venuje s osobitným zreteľom na oblasť kybernetickej bezpečnosti, etiky a informačnej spoločnosti.

Recenzovali

prof. ThDr. Marian Šuráb, PhD., RKCMBF UK Bratislava

doc. ThDr. Radoslav Lojan, PhD., TF KU Košice

doc. ThDr. Patrik Maturkanič, PhD., VŠAPs Terezín

ThDr. Ing. Vladimír Thurzo, PhD., RKCMBF UK Bratislava

Prínosom je jednoznačne komplexný a ucelený prehľad poznatkov z oblasti príspevku rozvoja systémov umelej inteligencie a navrhnuté riešenia etických problémov AI do budúcnosti.

doc. ThDr. Radoslav Lojan, PhD.

Způsob, jakým autor zpracoval předložené téma, vypovídá o velké erudici. Nabízený text systematicky přináší aktuální a relevantní zjištění, která lze považovat za významná, jak pro teorii, tak následnou praxi.

doc. ThDr. Patrik Maturkanič, PhD.

Publikácia sa môže stať „vademékom“ pre ďalšie práce, ktoré sa budú venovať problematike umelej inteligencie.

prof. ThDr. Marian Šuráb, PhD.

Možno povedať, že publikácia je jedinečná a novátorská najmä preto, že téma bola spracovaná práve na bohosloveckej fakulte. Jednak ide o oblasť, ktorej sa dosiaľ nikto nevenoval, ale najmä vzhľadom na jej interdisciplinárne, či skôr multidisciplinárne uchopenie... Vzhľadom na obsah, rozsah spracovania a najmä interdisciplinaritu som presvedčený, že ako monografia bude mať toto dielo svoje miesto na knižnom trhu.

ThDr. Ing. Vladimír Thurzo, PhD.

Abstrakt

Publikácia predstavuje určité nóvum, snažiac sa o inovatívny prístup v nazeraní na problematiku umelej inteligencie v kontexte kresťanského svetonázoru a etických dôsledkov limitov a rizík, s ktorými sa táto atraktívna oblasť moderného technologického sveta v súčasnosti potýka: interdisciplinárne uchopenie fenoménu umelej inteligencie, dôsledná analýza limitov i rizík a z nej prameniace etické závery, návrh základnej štruktúry všeobecných i špecifických etických princípov a zásad, vyjadrenie diskutabilných faktorov i niektorých perspektív všeobecnej umelej inteligencie smerujúcej k superinteligencii a zdôvodnenie jasného postoja k jej porovnávaniu s ľudskou bytosťou – to všetko tvorí vklad do prebiehajúceho celosvetového etického diskurzu, ktorý tým viac naberá na dôležitosti, čím viac sa technológie umelej inteligencie rozvíjajú a ich sofistikované implementácie sa do reálneho života masívne zavádzajú, ovplyvňujúc tak životy miliónov ľudí.

Publikácia vychádza z autorovej dizertačnej práce *Etické výzvy a morálne aspekty súčasných systémov umelej inteligencie**, pričom pôvodný text bol upravený a podstatne rozšírený o ďalší výskum v oblasti všeobecnej umelej inteligencie.

---

* ŠANTAVÝ, P. *Etické výzvy a morálne aspekty súčasných systémov umelej inteligencie* [dizertačná práca]. [on-line]. Bratislava: RKCMBF UK, 2022. Dostupné na internete: <https://peter.santavy.cloud/data/uploads/docs/eticke_vyzvy_a_moralne_aspekty_sucasnych_systemov_umelej_inteligencie.pdf>

Venované rodičom, ktorí sa vydania tejto knihy už nedožili.

Motto

„Umelá inteligencia“ (artificial intelligence - AI) prináša a bude čoraz viac prinášať veľké zmeny do života ľudí. Ponúka totiž mimoriadne možnosti...

...teraz teda viac ako inokedy musíme zaistiť taký prístup, v ktorom sa rozvoj AI nesústreďuje len na technológiu, ale berie ohľad na dobro ľudstva a životného prostredia, na náš spoločný domov a jeho ľudských obyvateľov, ktorí sú navzájom nerozlučne prepojení.

Výzva na etiku v oblasti umelej inteligencie

Obsah

Table of Contents

Zoznam obrázkov

Zoznam skratiek a symbolov

AI	artificial intelligence (umelá inteligencia)
ANI	artificial narrow intelligence (narrow AI)
AGI	artificial general intelligence (general AI)
ASI	artificial super intelligence
CNN	konvolučné neurónové siete (convolutional neural networks)
ConvNets	konvolučné neurónové siete (convolutional neural networks)
GEB	zaužívaný akronym pre knihu *Gödel, Escher, Bach: an Eternal Golden Braid*, ktorú v roku 1979 napísal Douglas Hofstadter
IKT	informačné a komunikačné technológie
IS	informačná spoločnosť
LAWs/AWs	lethal autonomous weapons – smrtiace autonómne zbraňové systémy
NhRP	Non-human Rights Project

Predslov (alebo čoho sa Douglas Hofstadter obáva)

Keď v roku 1979 Douglas Hofstadter[1] napísal knihu *Gödel, Escher, Bach: an Eternal Golden Braid*[2] (známu tiež pod skratkou GEB), určite nečakal, koľko súčasných vedcov a pedagógov venujúcich sa umelej inteligencii privedie k tejto problematike práve jeho kniha[3].

Možno i preto bol Douglas Hofstadter v roku 2014 pozvaný do centrály Google, aby tam prednášal vybraným vedcom z oblasti umelej inteligencie. A keďže viacerí z nich na GEB vyrastali, tridsaťpäť rokov od jej napísania bol dosť dlhý čas na rekapituláciu vízií, ktorú nemohol urobiť nik povolanejší, než sám autor.

Zvedavému auditóriu miesto očakávaného nadšenia, predstavených vízií, povzbudenia, či rád do vedeckej práce legendárny Douglas Hofstadter prezentoval obavy a zdesenie.

V roku 1979 bola téma umelej inteligencie vzrušujúcou, no veľmi vzdialenou budúcnosťou; vzdialenou až do tej miery, že akákoľvek jej možná realizácia a riziká boli brané ako niečo, čo sa týka budúcich generácií. Douglas Hofstadter – narozdiel od viacerých iných odborníkov – veril v realizovateľnosť umelej inteligencie napodobňujúcej inteligenciu človeka, no v jeho podaní išlo viac o víziu, než o projekt, ktorý by sa mohol v najbližších

---

1 Douglas Richard Hofstadter je americký filozof, spisovateľ, vysokoškolský pedagóg, informatik a fyzik, ktorý sa zaoberá interdisciplinárnymi témami, ako napr. vedomie, preklad, tvorivosť a porovnávanie ľudského rozumu a umelej inteligencie. Je legendou v oblasti umelej inteligencie.
Douglas Hofstadter [on-line]. [cit. 30. júla 2020].
Dostupné na internete: <https://en.wikipedia.org/wiki/Douglas_Hofstadter>

2 Jednoducho povedané, GEB je súhrnom intelektuálnych záujmov Douglasa Hofstadtera (matematika, umenie, hudba, reč, humor a slovné hry), ktoré sú premietnuté do fundamentálnej otázky, ako sa inteligencia, uvedomenie a pocity sebauvedomenia, vlastné ľudským bytostiam, môžu tak zásadne vynoriť z neinteligentného, nevedomého substrátu biologických buniek. A tiež ako k inteligencii a sebauvedomeniu môžu eventuálne dospieť počítače.
HOFSTADTER, D. *Gödel, Escher, Bach: an Eternal Golden Braid*. New York: Basic Books, 1979.

3 Za všetkých môžeme spomenúť Dr. Melanie Mitchell, bývalú doktorandku a spolupracovníčku Douglasa Hofstadtera na University of Michigan, ktorá v súčasnosti pôsobí ako profesorka počítačových vied na Portland State University. O svojej ceste k AI a podobnej skúsenosti viacerých vedcov z rôznych tímov z Google AI píše vo svojej knihe *Artificial Intelligence*.
MITCHELL, M. *Artificial Intelligence.* Farrar, Straus and Giroux. 2019, s. 5-6.

desaťročiach reálne uskutočňovať.[4]

V závere GEB Hofstadter uvádza desať otázok a špekulácií o umelej inteligencii.

Jednou z nich bola i otázka: Bude existovať šachový program, ktorý by porazil *kohokoľvek*?

Hofstadter odpovedá: nie, nebude existovať (maximálne tak porazí dobrého amatéra). Na porazenie šachového majstra by muselo ísť o počítačový systém s univerzálnou inteligenciou.[5]

Hofstadter vychádzal z analýz svojho priateľa Eliota Hearsta, šachového šampióna a profesora psychológie. Hearst sa venoval rozdielom medzi ľudským šachovým expertom a počítačovým šachovým programom, tvrdiac, že zatiaľčo sa človek rozhoduje na základe rozpoznávania vzorov a abstrahovania, program koná na základe masívneho vyhľadávania budúcich ťahov hrubou výpočtovou silou[6]. Podľa Hearsta bez všeobecnej inteligencie a jej zodpovedajúcej úrovni abstrakcie konceptov nie je možné dosiahnuť úroveň šachového veľmajstra.

Osemdesiate a deväťdesiate roky minulého storočia však priniesli veľký pokrok nielen vo výpočtovom výkone počítačov, ale i v šachových programoch pracujúcich v princípe stále len *hrubou silou*. Neustále zlepšovanie vyústilo v roku 1997 do svetoznámeho duelu medzi Deep Blue machine od IBM a úradujúcim svetovým šampiónom Garry Kasparovom. Deep Blue v zápase šiestich hier Kasparova porazil.[7]

Šachové majstrovstvo, ktoré bolo kedysi považované za vrchol ľudskej inteligencie,

---

4 Sám sa vyjadroval, že realizácia človeku podobnej umelej inteligencie je vzdialená na *sto Nobelových cien* (citované v ROTA G.-C. *Indiscrete Thoughts*. Boston: Berkhäuser, 1997, s. 22).

5 HOFSTADTER, *Gödel, Escher, Bach: an Eternal Golden Braid*, s. 678.

6 Je to zjednodušene a skôr principiálne vyjadrené, keďže už od roku 1949 Arthur Samuel, inžinier v IBM a jeden z priekopníkov v oblasti umelej inteligencie (práve on šíril termín strojové učenie / machine learning), vyvíjal program pre hru dámy, v rámci ktorého boli položené viaceré základy z budúcich metód umelej inteligencie, napr. strom hry, vyhodnocovacie funkcie, učenie sa samou hrou s načrtnutými základmi pre učenie posilňovaním (reinforcement learning). Viac napríklad v publikácii: SAMUEL, A. L. *Some Studies in Machine Learning Using the Game of Checkers*. In: *IBM Journal of Research and Development.* 1959, č. 3, s. 210-229.

7 *Deep Blue versus Garry Kasparov* [on-line]. [cit. 30. júla 2020]. Dostupné na internete: <https://en.wikipedia.org/wiki/Deep_Blue_versus_Garry_Kasparov>

podľahlo postupom využívajúcim hrubú výpočtovú silu...[8]

Pre Douglasa Hofstadtera bola hudba *baštou človečenstva* – jednoducho doménou i výkladnou skriňou ľudského ducha. Preto – po zlyhaní šachovej vízie – prichádza na rad ďalšia z desiatich otázok, ktorými kedysi Douglas Hofstadter zavŕšil GEB: Bude vedieť počítač skomponovať *nádhernú* hudbu?

Hofstadter predpovedá: áno, no nebude to čoskoro. Ďalej vysvetľuje: hudba je jazykom emócií, a pokiaľ programy nedisponujú emóciami podobne komplexnými, ako sú ľudské, nie je možné, aby program skomponoval čokoľvek nádherné, či krásne. Môžu existovať falzifikáty a povrchné imitácie vychádzajúce z poznania a syntaxe už vytvorenej hudby, no to určite nestačí na tvorivú činnosť, ktorej ovocím je krásna hudba. Pre samotného Douglasa Hofstadtera bola táto úvaha tak dôležitá, že – ako sám hovorieval – na toto

---

8 MITCHELL, *Artificial Intelligence*, s. 8.
Treba dodať, že šach nie je vhodným základom pre meranie schopností umelej inteligencie, či už vzhľadom na reálne vypočítateľný počet možných ťahov alebo algoritmizáciu šachových pravidiel (Deep Blue – narozdiel napr. od Samuelovho programu na hru dámy – neobsahoval žiadne pokročilejšie metódy umelej inteligencie). Oveľa náročnejšou skúškou je čínska abstraktná strategická hra GO, ktorá je oveľa zložitejšia než šach. GO umožňuje vytvoriť $2x10^{170}$ pozícií, čo je oveľa viac, než je atómov v známom vesmíre (10^{80}). Jednoduchá algoritmizácia a hrubá výpočtová sila tu zlyhávajú.
I napriek tomu v rokoch 2015 – 2017 umelá inteligencia AlphaGo od Googlu porazila viacerých najlepších hráčov sveta. Jej predposledná verzia AlphaGo Zero, využívajúc pokroky v AI, porazila pôvodnú AlphaGo 100:0 a najnovšia verzia AlphaZero je považovaná za najlepšieho hráča Go a pravdepodobne i šachu (narozdiel od Deep BLue však všetky verzie AlphaGo pokročilé metódy umelej inteligencie využívajú – ide o použitie známeho vyhľadávania pomocou stromu Monte-Carlo a GPT, t.j. Generative Pre-Trained Transformer).
AlphaGo [on-line]. [cit. 30. júla 2020].
Dostupné na internete: <https://en.wikipedia.org/wiki/AlphaGo>
O AlphaZero sa niekedy hovorí ako o prekročení slabých umelých inteligencií (o slabej a silnej umelej inteligencii budeme pojednávať v ďalšom texte).
TUROŇ, J. *Alpha Zero: soumrak slabých umělých inteligencí* [on-line]. [cit. 30. júla 2020].
Dostupné na internete: <https://www.osel.cz/9702-alpha-zero-soumrak-slabych-umelych-inteligenci.html>
Aj my sa skôr prikláňame k názoru prof. Mitchellovej, prof. Russella a mnohých iných odborníkov na umelú inteligenciu, tvrdiac, že ani pri takých systémoch, ako sú AlphaGo, GPT-3 a pod. sa stále nedá hovoriť o prekračovaní hraníc slabej umelej inteligencie, aj keď treba povedať, že seriózny vývoj v tejto oblasti neustále prebieha (napr. projekt Alexandria, COMET, atď.).

tvrdenie by vsadil aj život.[9]

I toto tvrdenie ohľadom umelej inteligencie však bolo v priebehu ďalších desaťročí otrasené. V deväťdesiatych rokoch sa Hofstadter stretol s programom EMI (Experiments in Musical Intelligence), ktorý napísal hudobník, skladateľ a profesor hudby David Cope. Pôvodným zámerom pre napísanie EMI bolo vytvoriť softvérovú podporu pre Copeho komponovanie hudby, ktorá by bola schopná automaticky vytvárať časti kompozície v osobitnom Copeho štýle. Napriek tomu sa EMI preslávila produkovaním diel v štýle klasických skladateľov, akými boli Bach a Chopin.

EMI *komponovala* využívajúc veľkú množinu pravidiel, ktoré vytvoril Cope a ktoré boli zamerané na zachytenie všeobecnej syntaxe komponovania. Tieto pravidlá boli následne aplikované na mnohé príklady z konkrétnych diel skladateľov s cieľom vytvoriť nové dielko v konkrétnom štýle konkrétneho skladateľa.

Sám Hofstadter to na stretnutí v centrále Google emotívne opisuje takto: „Sadol som si k svojmu klavíru a hral som jednu z mazuriek vytvorených programom EMI podľa štýlu Chopinových mazuriek. Neznelo to presne ako Chopin, avšak znelo to veľmi podobne ako Chopin a ako súvislá hudba, takže som sa cítil *hlboko* znepokojený.“[10]

Toto veľké znepokojenie u Hofstadtera vychádzalo z jeho prežívania hudby ako emočného dotyku s osobou, ktorá ju skomponovala. Hudbu vnímal ako možnosť priameho prístupu do duše skladateľa, a preto pre neho na svete neexistovalo nič ľudskejšieho ako vyjadrenie sa hudbou. Myšlienka, že povrchná manipulácia so vzormi môže priniesť veci, ktoré znejú, akoby vychádzali z ľudského srdca, bola pre neho veľmi znepokojujúca a jeho úvahy to úplne rozbilo.[11]

Toto znepokojenie bolo ešte umocnené, keď následne požiadal osadenstvo prestížnej Eastman School of Music v Rochesteri, v New Yorku, aby porovnali konkrétnu mazurku od Chopina a mazurku skomponovanú programom EMI. Hofstadter bol šokovaný, keď mnohí z fakulty označili kompozíciu od EMI ako pravé dielo Chopina a to skutočné Chopinovo dielo ako menej dokonalý výtvor EMI.[12]

---

9 HOFSTADTER, *Gödel, Escher, Bach: an Eternal Golden Braid*, s. 676.

10 MITCHELL, *Artificial Intelligence*, s. 9.

11 Por. MITCHELL, *Artificial Intelligence*, s. 9.

12 HOFSTADTER, D. *Staring Emmy Straight in the Eye – and Doing My Best Not to Flinch*. In: DARTNELL, T. *Creativity, Cognition, and Knowledge*. Westport, Conn.: Praeger, 2002, s. 67-100.

Douglas Hofstadter túto skúsenosť zaujímavo hodnotí: „Bol som zhrozený z EMI, nenávidel som ju a cítil som sa ňou extrémne ohrozený. EMI hrozilo zničením toho, čo som si najviac cenil na ľudskosti. Myslím, že EMI bolo najtypickejším príkladom obáv, ktoré som mal z umelej inteligencie.“[13]

Medzi spoločnosťami, ktoré sa venujú výskumu a vývoju umelej inteligencie, zaberá (aj vzhľadom na svoje akvizície technologických start-up-ov) Google vedúce postavenie.[14] O to viac je pre Hofstadtera znepokojujúce vedomie, že práve Google si osvojil víziu Raya Kurzweila[15] o *singularite*[16], v ktorej umelá inteligencia so schopnosťou učiť sa, zlepšovať sa a samostatne sa vyvíjať rýchlo dosiahne a následne prevýši inteligenciu človeka.

Hoci Hofstadter – a nielen on – vážne pochybuje o predpoklade singularity v oblasti umelej inteligencie, vidiac pokroky, ktoré sa v oblasti vývoja umelej inteligencie za posledné desaťročia dosiahli, pociťuje obavy a rozrušenie z možného naplnenia Kurzweilovej predikcie. Na stretnutí v centrále Google to komentuje takto: „Bol som vydesený týmto scenárom. Veľmi skeptický, no v rovnakom čase, mysliac si, že i keď predpokladaný

---

13 MITCHELL, *Artificial Intelligence*, s. 10.
V kontexte pokročilej umelej inteligencie sa EMI budeme ešte venovať v 5. kapitole.

14 Výskum, rozvoj a aplikácia umelej inteligencie má v portfóliu Google (resp. v súčasnosti Alphabet) veľmi široké zastúpenie: autonómne vozidlá, rozpoznávanie reči a inteligentná strojová asistencia, pochopenie prirodzeného jazyka a metódy prekladu, počítačovo generované umenie, riešenie logických hier (AlphaGO až AlphaZero), inteligentné humanoidné roboty a pod.

15 Ray Kurzweil je svetoznámy vynálezca a kontroverzný futurista, ktorý predstavil myšlienku singularity umelej inteligencie. Google zamestnal Kurweila, aby pomohol túto víziu zrealizovať.
Por. MITCHELL, *Artificial Intelligence*, s. 4.

16 Pojem singularita sa v mnohých oblastiach matematiky a fyziky používa na vyjadrenie stavu, ktorý je zvláštny, nedefinovaný, odchyľujúci sa z trendu, či množiny, stav prekračujúci očakávajúce parametre, nekonečný, či v daných podmienkach neriešiteľný (napr. aktuálne fyzikálne modely, ktorými nedokážeme popísať singulárne body, akými sú napr. Veľký tresk a čierne diery).
Pod termínom technologická singularita sa myslí teoretický bod vo vývoji vedeckej civilizácie (znalostnej spoločnosti), v ktorom sa technologický pokrok zrýchli do nekonečna a prevýši všetky predpovede.
Singularita [on-line]. [cit. 3. augusta 2020].
Dostupné na internete: <https://sk.wikipedia.org/wiki/Singularita>
Singularitou v umelej inteligencii je myslený stav, ktorý nastane, ak umelá inteligencia so schopnosťou učiť sa, zlepšovať sa a samostatne sa vyvíjať rýchlo dosiahne a následne prevýši inteligenciu človeka. Teda situácia, keď sa počítačové systémy stanú inteligentnejšími než ľudia.
Por. MITCHELL, *Artificial Intelligence*, s. 10.

časový harmonogram je nereálny, možno majú pravdu. A potom budeme úplne zaskočení. Budeme si myslieť, že sa nič nedeje a zrazu, skôr než si to uvedomíme, budú počítače múdrejšie než my.“ A keď sa to stane, „budeme nahradení, staneme sa reliktom. Zostaneme v prachu.“ Končiac svoj príhovor dodáva: „Možno sa to stane, ale nechcem, aby sa to stalo priskoro. Nechcem, aby moje deti zostali v prachu.“[17]

Priamo na inžinierov, vývojárov a vedcov z Google sa v závere Douglas Hofstadter obracia, hovoriac: „Považujem to za veľmi hrozivé, veľmi znepokojujúce, veľmi smutné a tiež považujem za hrozné, desivé, bizarné, bezradné i zarážajúce to, že ľudia sa slepo a akoby v ošiali hrnú do vytvárania týchto vecí.“[18]

Sú obavy Douglasa Hofstadtera opodstatnené? Alebo sú len ovocím prekročenia jeho vízií o rozvoji umelej inteligencie spôsobom, ktorý svojho času pre neho, ale i pre druhých nebolo možné predikovať? V tomto diele sa pokúsime uchopiť, pomenovať a možno i definovať viaceré aspekty z oblasti umelej inteligencie, aby sme si mohli ozrejmiť základné hodnoty a postoje, vďaka ktorým i systémy umelej inteligencie môžu slúžiť nášmu proaktívnemu postoju k rozvoju sveta vo svetle evanjelia a pozvania, ktoré sme od Pána dostali.[19]

17 MITCHELL, *Artificial Intelligence*, s. 10-11.

18 MITCHELL, *Artificial Intelligence*, s. 10-11.

19 Autor diela je zároveň katolíckym kňazom.

Uvedenie do problematiky

Prežívame dejinný úsek, ktorý sa hrdí takými prívlastkami ako digitálna disrupcia, rodiaca sa informačná a znalostná spoločnosť, zmena paradigmy, permanentná technologická revolúcia, atď. V tomto kontexte takmer každá technologická novinka sa pýši dodatkom, že v jej útrobách je implementovaná umelá inteligencia. Nájdeme a vieme vymenovať veľa oblastí, v ktorých prvky umelej inteligencie zohrávajú čoraz dôležitejšiu úlohu: v lekárskej diagnostike a liečbe, v riadení technologických procesov, predikcii vývoja, spracovaní obrazu, analýze a syntéze reči, kybernetickej bezpečnosti[20], doprave, zábave, komunikačnej technike, genetickom výskume, jadrovej fyzike, astrofyzike, atď.

Javí sa, že umelá inteligencia nie je len *buzzword*, ktorý má pomôcť technologickým firmám presadiť sa a zarobiť, či okrášliť stratégie lídrov dnešného sveta, ale ide o reálny koncept a integrálnu súčasť technologickej budúcnosti našej civilizácie. Systémy umelej inteligencie vo viacerých oblastiach posúvajú úroveň poznania a skúmanie míľovými krokmi vpred. Dokážu byť extrémne nápomocné pri záchrane ľudských životov, ochrane zdravia i zaradení sa do reálneho života po ťažkých chorobách a úrazoch. Stávajú sa takmer nepostrádateľnými asistentmi v bežnom živote, v doprave, komunikácii, zábave. Prinášajú extrémny posun vpred pri analýze a spracovaní vedeckých dát. Pridávajú ďalší stupeň ochrany pri bezpečnostných systémoch, v obrane, v boji so zločinom a pod. Takmer v každej oblasti ľudskej činnosti a života človeka nájdeme niečo, v čom sa použitie prvkov umelej inteligencie stáva veľmi osožným.

Využitie umelej inteligencie má však aj svoju temnú stranu a riziká: jednou z privilegovaných oblastí vývoja umelej inteligencie sú autonómne zbraňové systémy a samostatné nasadenie v boji; využitie prvkov umelej inteligencie v kybernetickej kriminalite je už teraz nočnou morou informačnej bezpečnosti; zneužitie umelej inteligencie pri rozpoznávaní tvárí, komplexnom monitoringu a kategorizácii ľudí, detekcii a predikcii vývoja v spoločnosti je svätým grálom akéhokoľvek autoritárskeho režimu; vplyv sociálnych médií poháňaných aktuálne nastavenými, či zneužitými algoritmami umelej inteligencie na psychiku človeka a rozvoj spoločnosti sa už teraz podľa niektorých odborníkov javí ako katastrofálny; vytváranie psychologických, zdravotných,

20 Sám autor k veľkej spokojnosti využíva algoritmy umelej inteligencie v oblasti kybernetickej bezpečnosti a ochrany dát.

sociologických, či iných profilov ľudí spájaním informácií z verejných zdrojov, sociálnych sietí a metadát dokáže vďaka umelej inteligencii vytvoriť nekompromisnú verejnú sondu do ľudskej duše a bezprecedentne odhaliť súkromie človeka[21]; zneužitie falošnej identity, či vytvorenie falošných informácií o človeku[22] bravúrne natrénovaným systémom umelej inteligencie môže danú osobu priviesť k totálnemu spoločenskému i osobnému kolapsu...

V kontexte vyššie uvedeného – ak také osobnosti, ako napr. Ray Kurzweil, predikujú umelej inteligencii schopnosť konkurovať, ba až nahradiť ľudskú bytosť a iní, ako napr. Douglas Hofstadter, vidia reálne riziká vyplývajúce z potenciálu napredovania v jej vývoji, **pridávame sa k hlasom, ktoré volajú po etickom zhodnotení, zvážení morálnych aspektov a definovaní pravidiel pre vývoj, používanie i samotné fungovanie systémov umelej inteligencie.**

Uvedomujeme si, že až na pár výnimiek, sa v súčasnosti umelou inteligenciu nazývajú informačné systémy, ktoré majú pramálo spoločné s víziou Kurzweila a obavami Hofstadtera[23], t.j. s víziou človeku konkurujúcej *skutočnej* umelej inteligencie. Nech však ide o akýkoľvek stupeň, či schopnosti týchto systémov, treba k nim zodpovedne pristupovať, aby sme v spoločnosti nielenže znovu nemuseli boľavo objavovať to známe *dobrý sluha a zlý pán*, ale aj nakoniec nezistili, že z umelej inteligencie dobrého sluhu vo svojej podstate ani nedokážeme reálne vytvoriť[24].

---

21 Systémy umelej inteligencie sú na základe analyzovaných dát a metadát častokrát schopné vytvoriť psychologický profil osoby, odhaliť jeho sexuálnu orientáciu, detekovať psychické problémy a pod.

22 Napr. vytvorenie falošnej videokonferencie (jedna z foriem tzv. DeepFake), na ktorej sa verejný činiteľ priznáva k veciam, ktoré nevykonal; alebo vytvorenie falošného videa, na ktorom človek s vysokým morálnym kreditom v spoločnosti koná nemravné veci; a pod.

23 Vo väčšine implementácií skôr ide o IA (intelligent assistance) a nie o AI (artificial intelligence).

24 Toto riziko sa reálne zvažuje pri debatách o všeobecnej umelej inteligencii (rôzne formy superinteligencie) nielen v rovine matematických teorémov popisujúcich komplexnosť takých systémov (napr. Riceov teorém), ale – ako prirodzený a nutný rozmer zodpovedného využitia superinteligencie – aj v rovine etickej a filozofickej.
CHOI, Q. CH. *Superintelligent AI May Be Impossible to Control; That's the Good News* [on-line]. [cit. 26. januára 2021].
Dostupné na internete: <https://spectrum.ieee.org/tech-talk/robotics/artificial-intelligence/super-artificialintelligence>
ALFONSEC, M., CEBRIAN, M. et al. *Superintelligence Cannot be Contained: Lessons from Computability Theory* [on-line]. [cit. 26. januára 2021].
Dostupné na internete: <https://jair.org/index.php/jair/article/view/12202>

Pri úvahe o dobrom sluhovi a zlom pánovi sme si podvedome dovolili ignorovať diskusiu o uznaní práv osoby pre pokročilé systémy umelej inteligencie, tzv. superinteligencie. Je to však diskusia, ktorá sa určitým spôsobom dotýka najhlbšieho jadra Hofstadterových obáv ohľadom umelej inteligencie a s ňou súvisiacej fiktívnej alebo reálnej singularity. Ak sa nad tým zamyslíme – podobne, ako to po opisovanej prednáške v centrále Google urobila prof. Melanie Mitchellová[25] – Hofstadterove obavy nie sú ničím novým. Svet, ktorý nerozumie technológiám umelej inteligencie a je naviac živený katastrofickými scenármi science fiction, je týchto obáv už niekoľko desaťročí plný.

Novými sú skôr dva fenomény:

- obavy spojené s rizikami pokročilej umelej inteligencie čoraz viac vyjadrujú ľudia, ktorí sú v tejto oblasti doma a častokrát patria k prvotriednym odborníkom, či technologickým vizionárom.[26] Netreba však obavy spájať len s víziou nadľudskej

25 MITCHELL, *Artificial Intelligence*, s. 11-12.

26 Už v roku 2014 svetoznámy fyzik Stephen Hawking vyhlásil: „Vývoj úplnej umelej inteligencie by mohol znamenať koniec ľudskej rasy.“ A vysvetľuje: „Ľudia, ktorí sú obmedzení pomalou biologickou evolúciou, nemohli by (s umelou inteligenciou) súťažiť a boli by nahradení.“
CELLAN-JONES, R. *Stephen Hawking Warns Artificial Intelligence Could End Mankind*. In: *BBC News.* [on-line]. 2014, 2. 12. [cit. 5. augusta 2020].
Dostupné na internete: <https://www.bbc.com/news/technology-30290540>
Elon Musk, vizionár a zakladateľ spoločností Tesla a SpaceX, v tom istom roku varuje pre umelou inteligenciou hovoriac, že ide pravdepodobne o „naše najväčšie existenciálne ohrozenie“ a že „umelou inteligenciou privoláme démona.“
MCFARLAND, M. *Elon Musk: „With Artificial Intelligence, We Are Summoning the Demon“*. In: *Washington Post*, 2014, 24. 10.
Elon Musk pokračoval v tejto zmysle aj vo svojich varovaniach z r. 2019 a 2021.
Elon Musk pritom v rámci Tesly vyvíja jednu z najpokročilejších umelých inteligencií pre autonómne vozidlá a v rámci svojich vizionárskych aktivít (Neuraling Corporation) vytvára rozhranie pre rozšírenie možností človeka, ako napr. prepojenie ľudského mozgu a umelej inteligencie pre dosiahnutie symbiózy, ktorú by sme poľahky mohli zaradiť do oblasti transhumanizmu.
Neuralink [on-line]. [cit. 5. augusta 2020].
Dostupné na internete: <https://en.wikipedia.org/wiki/Neuralink>
Bill Gates, spoluzakladateľ softvérového gigantu Microsoft, dopĺňa: „V tomto súhlasím s Elonom Muskom i viacerými ďalšími a nemôžem pochopiť, prečo niektorí ľudia nie sú znepokojení.“
Hi Reddit, I’m Bill Gates and I’m back for my third AMA. Ask me anything. In: *Reddit.* [on-line]. 2015, 28. 1. [cit. 7. augusta 2020].
Dostupné na internete:

umelej inteligencie – oveľa aktuálnejšie je riešenie mnohých výziev (medzi nimi i etických) spojených s prvkami umelej inteligencie, ktoré sú aktuálne vyvíjané a začínajú sa masovo používať naprieč rôznymi oblasťami modernej spoločnosti[27];

- riziká spojené s pokročilou umelou inteligenciou sú často spojené s disproporciou medzi vnímaním ľudstva a umelou inteligenciou[28], a tak nastavujú zrkadlo nášmu pohľadu na podstatu ľudského bytia[29]: kým sme, čo nás definuje, do akej miery sme redukovateľní na technologické vyjadrenie, aká je hodnota a podstata ľudského

---

<https://www.reddit.com/r/IAmA/comments/2tzjp7/hi_reddit_im_bill_gates_and_im_back_for_my_third/>
Významné varovanie je vyjadrené aj v knihe *Human Compatible* od v súčasnosti jednej z najväčších autorít v oblasti umelej inteligencie, prof. Stuarta Russella.
Vo svojej práci rieši i tzv. „zásadnú chybu v srdci umelej inteligencie, ktorá – ak nebude vyriešená – môže prerásť do katastrofy“ (pozri komentár od Iana Sample z Guardianu v rámci Books of the Year).
RUSSELL, S. *Human Compatible.* Penguin Books, 2020, s. III.
A Judea Pearl, počítačový vedec a filozof, autor jedného z princípov využívaných v systémoch umelej inteligencie i známej knihy *The Book of Why*, dodáva: „Human Compatible ma primälo osvojiť si Russelove starosti ohľadom našej schopnosti ovládať náš prichádzajúci výtvor – super inteligentné stroje. Keďže na rozdiel od alarmistov a futuristov mimo odboru je Russell vedúcou autoritou v oblasti umelej inteligencie.“
RUSSELL, *Human Compatible*, s. IV.

27 Viacerí významní autori (Mitchel Kapor, Rodney Brooks, Gary Marcus,...) hovoria: „áno, mali by sme sa ubezpečiť, že systémy umelej inteligencie sú bezpečné a nepoškodzujú ľudí, ale akékoľvek správy o nadľudskej AI sú veľmi prehnané.“
Por. MITCHELL, *Artificial Intelligence*, s. 13.

28 Vedomie, že mnohí vývojári systémov pokročilej umelej inteligencie podceňujú inteligenciu ľudí, nás poľahky môže priviesť k opačnému extrému – podceňovaniu výkonu, možností a dôvery či nádejí vkladaných do súčasných technológií umelej inteligencie.
Por. MITCHELL, *Artificial Intelligence*, s. 12-13.

29 Z tohto vlastne pramenia Hofstadterove obavy – či sa jeho predstava o ľudskom bytí a jeho hodnote nerozpadne do ním spomínaného *prachu*: nie je to o umelej inteligencii, ktorá sa stáva oveľa múdrejšou, príliš invazívnou, veľmi škodlivou alebo priveľmi použiteľnou. Je to zhrozenie z uvedomenia si, že inteligencia, kreativita, emócie a možno dokonca i vedomie by mohli byť príliš ľahko reprodukovateľné a vytvárané balíkom trikov, vďaka ktorým povrchný súbor algoritmov spojený s hrubou výpočtovou silou môže vysvetliť ľudského ducha.
A dodáva: „Ak by sa takéto mysle nekonečnej jemnosti a komplexnosti a emocionálnej hĺbky (myslí tým velikánov ľudstva) dali trivializovať malým čipom, zničilo by to môj pohľad na človečenstvo.“
MITCHELL, *Artificial Intelligence*, s. 11-12.

bytia, atď... Teda otázky skôr filozofické, antropologické a tiež i teologické[30], ktoré sú tým aktuálnejšie, čím viac sa posúvame vo vývoji systémov umelej inteligencie dopredu a čím viac rastie náš apetít po vytvorení umelej inteligencie schopnej konkurovať ľudskému bytiu[31].

V skutočnosti sa však nachádzame vo svete, ktorý presahuje akademické úvahy o potenciáli a rizikách pokročilej umelej inteligencie. Viaceré moderné armády vyvíjajú a koketujú s nasadením autonómnych bojových systémov (USA, Rusko, Čína, Izrael,...)[32], v niektorých krajinách sa pripravuje legislatíva pre autonómne vozidlá[33], prvá krajina uznala práva pre inteligentného robota[34], prvým systémom, ktorý ešte pred prepuknutím pandémie ochorenia Covid-19 informoval, že prichádza nová epidémia koronavírusu, bola kanadská lekárska umelá inteligencia[35], lieky, ktoré by mohli účinkovať voči SARS-CoV-2

30 A tým zároveň ide o aj o otázky etické, legislatívne, právne, sociologické, psychologické, medicínske,...

31 Účastníci Hofstadterovej prednášky v centrále Google boli prakticky všetko ľudia, ktorých snahou bolo naplnenie Kurzweilovej vízie o umelej inteligencii schopnej v konečnom dôsledku prevýšiť inteligenciu človeka.
Por. MITCHELL, *Artificial Intelligence*, s. 12-13.

32 Konkrétnu citáciu neuvádzame, pretože v tejto oblasti je neustále prichádzajúcich správ také množstvo a v takej odbornej granularite, že pre utvorenie uceleného obrazu by sme mohli uviesť niekoľko desiatok citácií...

33 Krajiny, v ktorých už prebieha testovanie autonómnych automobilov, postupne pripravujú svoju legislatívu. Kalifornia ako jeden z prvých federálnych štátov USA zakomponovala do svojho zákonníka o vozidlách (Vehicle code) oddiel 16.6, (Divison 16.6. Autonomous Vehicles [38750]), ktorý upravuje oblasť autonómnych vozidiel.
Autonómne vozidlá [on-line]. [cit. 6. augusta 2020].
Dostupné na internete: <http://akgunis.sk/autonomne-vozidla/>

34 28. októbra 2017 udelila Saudská Arábia na podujatí Future Investment Initiative conference občianstvo robotovi Sofii.
Robot Sophia speaks at Saudi Arabia's Future Investment Initiative [on-line]. [cit. 6. augusta 2020].
Dostupné na internete: <https://youtu.be/dMrX08PxUNY>
PAROULKOVÁ, V. *Lidská práva pro roboty? Evropská unie chystá právní statut tzv. umělé osoby* [on-line]. [cit. 6. augusta 2020].
Dostupné na internete: <https://plus.rozhlas.cz/lidska-prava-pro-roboty-evropska-unie-chysta-pravni-statut-tzv-umele-osoby-6598059>

35 MIHULKA, S. *První varování před koronavirem Wuhan poslala umělá inteligence BlueDot*. [on-line]. [cit. 6. augusta 2020].
Dostupné na internete: <https://www.osel.cz/11002-prvni-varovani-pred-koronavirem-wuhan-poslala-

navrhla taktiež umelá inteligencia[36], začínajú sa realizovať lekárske operácie vedené systémami umelej inteligencie, japonská lekárska umelá inteligencia s veľkou presnosťou a prekračujúc schopnosti najlepších špecialistov detekuje vzácne typy rakoviny, umelá inteligencia sa podieľa na špičkových fyzikálnych výskumoch[37], prakticky celá špička firiem špecializujúcich sa na kybernetickú bezpečnosť masívne implementuje prvky umelej inteligencie do svojich riešení[38], pokroky v rozpoznávaní reči v spojení s ďalšími prvkami umelej inteligencie umožnili vznik inteligentných asistentov[39], systémy spracovania a rozpoznávania obrazu kategorizujú, titulkujú, filtrujú, či inak spracúvajú snímky a obrazový materiál v takmer všetkých cloudových riešeniach a koniec koncov primitívnejšie systémy umelej inteligencie sú súčasťou prakticky každého spracovania obrazu, navrhovania trás v mapových systémoch a marketingových nástrojov v elektronických médiách, mobilných zariadeniach a e-shopoch...

V kontexte informačnej spoločnosti sa teda už nebavíme o tom, či a kedy prvky umelej inteligencie nasadiť – veď sú tu medzi nami a ich nasadenie sa neustále rozširuje – ale ako, to znamená za akých podmienok, pre aké ciele, akým spôsobom a s akými dôsledkami by mala byť umelá inteligencia súčasťou nášho sveta. Súčasťou sveta, v ktorom má človeku a spoločnosti slúžiť (vizionár by povedal *koexistovať*). A preto ľudia i spoločnosť majú mať jasno v tom, ako vyzerá etický návrh,

---

umela-inteligence-bluedot.html>

36 BECK, B. R. et al. *Predicting commercially available antiviral drugs that may act on the novel coronavirus (2019-nCoV), Wuhan, China through a drug-target interaction deep learning model*. In: *bioRxiv.* [online]. 2020, s. 2020.01.31.929547. [cit. 6. augusta 2020]. DOI: 10.1101/2020.01.31.929547. Dostupné na internete: <https://www.biorxiv.org/content/10.1101/2020.01.31.929547v1>

37 Napr. HE, S. et al. *Learning to predict the cosmological structure formation*. In: *Proceedings of the National Academy of Sciences.* [online]. 2019, roč. 116, č. 28, s. 13825. [cit. 6. augusta 2020]. DOI: 10.1073/pnas.1821458116. Dostupné na internete: <https://www.pnas.org/content/116/28/13825> SUMMER, T. *The first AI universe sim is fast and accurate—and its creators don't know how it works*. [on-line]. [cit. 6. augusta 2020]. Dostupné na internete: <https://phys.org/pdf480780725.pdf>

38 Miesto uvedenia všetkých významných implementátorov umelej inteligencie (CheckPoint, Kaspersky, Sophos, Cisco, Eset, IBM, Fortinet,...) v oblasti kybernetickej bezpečnosti by sme oveľa ľahšie vymenovali tých, ktorí tak ešte nekonajú.

39 Spomeňme Siri od Apple, Cortanu od Microsoftu, Now a Assistant z dielne Google, Alexu od Amazonu... *Virtual assistant* [on-line]. [cit. 6. augusta 2020]. Dostupné na internete: <https://en.wikipedia.org/wiki/Virtual_assistant>

realizácia a využívanie systémov umelej inteligencie, pričom samotná umelá inteligencia v svojom autonómnom a adaptívnom fungovaní „dokáže“ etické hodnoty a nastavené *morálne normy* dodržať[40].

V súčasnosti preto silnejú hlasy vyjadrujúce potrebu skúmať, navrhnúť, prijať a realizovať etický rámec vývoja, používania a fungovania umelej inteligencie – či už ide o oblasť vedeckého bádania[41] alebo vývoja a realizácie[42]. Od petícií a otvorených listov adresovaných OSN i niektorým vládam badať prechod k snahe systematicky túto oblasť podchytiť a preskúmať na národnej i medzinárodnej úrovni[43].

---

40 Viac krát sa stalo, že systémy s umelou inteligenciou museli byť vypnuté, lebo po uvedení do prevádzky sa vyvíjali nesprávnym smerom. Napr.:
KRAFT, A. *Microsoft shuts down AI chatbot after it turned into a Nazi*. [on-line]. [cit. 6. augusta 2020]. Dostupné na internete: <https://www.cbsnews.com/news/microsoft-shuts-down-ai-chatbot-after-it-turned-into-racist-nazi/>
HAMILTON, I. A. *Amazon built an AI tool to hire people but had to shut it down because it was discriminating against women*. [on-line]. [cit. 6. augusta 2020].
Dostupné na internete: <https://www.businessinsider.com/amazon-built-ai-to-hire-people-discriminated-against-women-2018-10>
GRIFFIN, A. *Facebook's artificial intelligence robots shut down after they start talking to each other in their own language*. [on-line]. [cit. 6. augusta 2020].
Dostupné na internete: <https://www.independent.co.uk/life-style/gadgets-and-tech/news/facebook-artificial-intelligence-ai-chatbot-new-language-research-openai-google-a7869706.html>

41 Viaceré iniciatívy poukazujúce na riziká a vyzývajúce k prijatiu potrebných pravidiel. Prehľad viacerých otvorených listov a petícií je uvedený na stránke:
Compilation of open letters against autonomous weapons. [on-line]. [cit. 19. augusta 2020].
Dostupné na internete: <https://autonomousweapons.org/compilation-of-open-letters-against-autonomous-weapons/>

42 Angažovanosť vyúsťujúca v kontexte aktuálnej politickej a spoločenskej situácie do viacerých aktivít vývojárov, napr. protest softvérovej inžinierky Laury Nolanovej proti zapojeniu Google do projektu Maven.
MCDONALD, H. *Ex-Google worker fears 'killer robots' could cause mass atrocities*. [on-line]. [cit. 19. augusta 2020].
Dostupné na internete: <https://www.theguardian.com/technology/2019/sep/15/ex-google-worker-fears-killer-robots-cause-mass-atrocities>
V kontexte udalostí rokov 2019-2021 – napr. nepokoje v USA a v Honkongu – sa viaceré korporácie pracujúce v oblasti aplikácie prvkov umelej inteligencie do bezpečnostných systémov rozhodlo obmedziť svoje aktivity, či zrušiť niektoré projekty a kontrakty.

43 Prehľad dokumentov do roku 2019 je sumarizovaný napríklad v tabuľke *Ethical guidelines for AI by country of issuer*, ktorá je súčasťou štúdie:

Naliehavosť tejto témy vyjadruje aj záujem, ktorý jej prikladá Katolícka cirkev. Po viacerých širšie koncipovaných aktivitách[44] Pápežská akadémia pre život vo februári 2020 zorganizovala konferenciu *renAIssance 2020*, ktorá sa venovala etike v oblasti umelej inteligencie. Jedným z výsledkov konferencie bolo aj podpísanie *Výzvy na etiku v umelej inteligencii* s cieľom „akcentovať etický prístup k umelej inteligencii a podporovať zmysel pre zodpovednosť medzi organizáciami, vládami a inštitúciami s cieľom vytvoriť budúcnosť, v ktorej digitálne inovácie a technologický pokrok slúžia ľudským schopnostiam a tvorivosti, a nie ich postupnému nahrádzaniu“[45]. Signatármi výzvy boli okrem Pápežskej akadémie pre život aj talianska vláda, FAO a také spoločnosti ako Microsoft, IBM a pod.

Zámerom konferencie bol nielen vstup do rozpravy o etických kritériách umelej inteligencie v jednotlivých oblastiach, ale aj diskusia so snahou o premostenie pohľadov jednotlivých krajín, firiem či záujmových skupín na oblasť etiky v problematike umelej inteligencie a s cieľom stanoviť konkrétne etické kritériá. Sumarizáciu tohoto úsilia vyjadruje vyššie spomenutý záverečný dokument, ktorého súčasťou je i formulovanie šiestich princípov pre použitie umelej inteligencie pre dobro človeka a bez strachu zo zneužitia.

Predložený náčrt problematiky umelej inteligencie v optike morálnych aspektov a etiky môže slúžiť ako uvedenie do tohto diela, v ktorej by sme ponajprv chceli predstaviť zjednodušenou formou základy umelej inteligencie v kapitole *Základy umelej inteligencie v skratke*. Pôvodne sme chceli uviesť, že znalý čitateľ môže túto kapitolu bez problémov

---

JOBIN, A., IENCA, M., VAYENA, E. *Artificial Intelligence: the global landscape of ethics guidelines*. [on-line]. [cit. 19. augusta 2020].
Dostupné na internete: <https://arxiv.org/pdf/1906.11668>

44 Napr. Hackaton 2018, ktorý sa však netýkal priamo umelej inteligencie, ale v duchu hesla *bridge the gap* skôr širšiemu využitiu informačných technológií na premostenie a podporu marginalizovaných, ktorí trpia napr. digitálnym rozdelením (digital divide) a pod. Viac o digital divide napr. v kapitole *Digitálne diferencovanie ľudstva (digital divide)* práce:
ŠANTAVÝ, P. *Analýza virtuálneho sveta v kontexte náboženskej formácie, pôsobenia a života Cirkvi* [licenciátska práca]. [on-line]. Bratislava: RKCMBF UK, 2017, s. 30-32. [cit. 19. augusta 2020].
Dostupné na internete:
<https://peter.santavy.cloud/data/uploads/docs/analyza_virtualneho_sveta_v_kontexte_nabozenskej_formacie_posobenia_a_zivota_cirkvi.pdf>

45 *Rome Call for AI Ethics*. [on-line]. [cit. 19. augusta 2020].
Dostupné na internete: <http://www.academyforlife.va/content/pav/en/events/intelligenza-artificiale.html>

preskočiť. Uvedomili sme si však, že mnohé, v tejto kapitole uvedené, postrehy a implikácie presahujú jednoduché vovedenie do problematiky AI a skôr sa priamo týkajú témy našej publikácie.

V kapitole *Limity a riziká súčasných systémov umelej inteligencie* sa pokúsime poukázať na limity a riziká existujúcich systémov AI, popísať technologické výzvy v oblasti bezpečnosti procesov umelej inteligencie a s tým súvisiacej kybernetickej bezpečnosti a akcentovať viaceré dôsledky v psychologickej, spoločenskej i vojenskej oblasti, ktorých zneužitie môže znamenať vážne ohrozenie pre jednotlivcov i spoločnosť.

V kapitole *Umelá inteligencia v optike etiky* sumarizujeme etické dôsledky limitov, rizík i možností existujúcich systémov AI, zmapujeme súčasné aktivity na poli etiky a regulácie umelej inteligencie a načrtneme aktuálny pohľad na etické otázky súvisejúce s vývojom a využívaním týchto systémov tak z pohľadu osobností a spoločností zaangažovaných v tejto oblasti, ako aj z pohľadu štátnych inštitúcií, medzinárodného spoločenstva a Cirkvi.

Kapitola *Navrhnuté riešenie etických problémov ANI* na základe analýzy súčasného stavu rozvoja systémov AI i predpokladaných možností ich ďalšieho vývoja a na základe porovnania i zhodnotenia existujúcich etických pravidiel a právnych noriem ponúkne pretavenie zistených skutočností do návrhu základnej štruktúry všeobecných i špecifických etických zásad, ktoré musia byť aplikované pri etickom návrhu, realizácii a využívaní systémov slabej umelej inteligencie (ANI).

I keď na poli vývoja uvedomelej umelej inteligencie[46] nebadať až tak veľké pokroky, ako by sa z rôznych mediálnych správ mohlo zdať, kapitola *Vízia silnej a všeobecnej umelej inteligencie* sa dotkne niektorých etických problémov aj v tejto vzrušujúcej a vizionárskej oblasti. Pravdu povediac – v tom rozsahu, v akom sa v tomto texte umelej inteligencii venujeme – debate o ľudskému bytiu podobnej umelej inteligencii, resp. superinteligencii sa v súčasnosti asi nedá vyhnúť.

Hlavným cieľom predkladanej publikácie je na základe analýzy aktuálneho stavu rozvoja systémov umelej inteligencie a existujúcich etických pravidiel snaha navrhnúť základnú štruktúru zásad, ktorá musí byť podchytená pri etickom návrhu, realizácii i využívaní akýchkoľvek systémov umelej inteligencie.

Sekundárnym aspektom našej snahy je v kontexte úsilia o pokorenie tzv.

46 Známej aj pod skratkou AGI (artificial general intelligence), ktorá však primárne prináleží označeniu silnej a všeobecnej umelej inteligencie.

koncepčných prelomov prekračujúcich možnosti súčasných systémov slabej, resp. úzkej umelej inteligencie (ANI) poukázať i na problematiku tzv. skutočnej umelej inteligencie, ktorá by podľa jej protagonistov mala byť dosiahnuteľná prostredníctvom technológií silnej a všeobecnej umelej inteligencie (AGI).

Z hľadiska metodológie používame viacero metód:

- analyticko-syntetickú metódu:
  - sumarizovanie aktuálneho stavu (technologického) rozvoja AI;
  - analýza známych limitov a rizík zlyhania i problematických dôsledkov nasadenia systémov AI;
  - analýza potencionálnych rizikových faktorov limitovanej všeobecnej umelej inteligencie;
  - analýza existujúcich etických pravidiel v oblasti AI.
- komparatívnu metódu:
  - porovnanie jednotlivých oblastí AI z pohľadu možnosti implementácie etických pravidiel;
  - porovnanie potencionálnych schopností sofistikovaných systémov AI s ľudským modelom správania, schopnosťou myslieť a vedomím.
- aplikáciu a návrh:
  - na základe analýzy súčasného stavu rozvoja systémov AI i predpokladaných možností ich ďalšieho vývoja a na základe porovnania i zhodnotenia existujúcich etických pravidiel predkladáme aplikáciu zistených skutočností do návrhu základnej štruktúry etických noriem, ktorá musí byť podchytená pri etickom návrhu, realizácii a využívaní akýchkoľvek systémov umelej inteligencie (ANI);
  - vklad do polemík o realizácii uvedomelej všeobecnej umelej inteligencie (AGI) na základe predloženej analýzy súčasných systémov AI a ich konfrontácie s ľudskou schopnosťou myslenia a sebauvedomenia;
  - návrh prístupu k vývoju všeobecnej umelej inteligencie v optike eliminovania jej negatívnych dôsledkov pre ľudské spoločenstvo.

V rámci aplikovaného výstupu v publikácii riešime analýzu aktuálneho stavu rozvoja systémov AI, komplexný popis limitov i rizík zlyhania a z toho prameniacich dôsledkov pre človeka i spoločnosť, a v neposlednom rade diskutujeme etické princípy a navrhované právne normy, resp. regulácie.

K základnému výskumu patrí porovnanie fenoménu uvedomelej umelej inteligencie a ľudského bytia v kontexte osoby a skutočnej schopnosti myslieť, definovanie limitovanej všeobecnej umelej inteligencie a rámcový návrh prístupu k jej vývoju v optike eliminovania negatívnych dôsledkov pre ľudské spoločenstvo, návrh interdisciplinárneho rámca a spôsobu prístupu k riešeniu problémov ANI, stanovenie základných etických princípov a noriem, ktoré musia byť základom pre etický návrh, realizáciu a využívanie systémov slabej umelej inteligencie vo všeobecnosti, a tiež formulácia špecifických odporúčaní a parciálnych usmernení pre osobitné oblasti nasadenia systémov AI.

1. Základy umelej inteligencie v skratke

Umelá inteligencia je ako Colombova žena, nikto ju nevidel a všetci o nej hovoria.[47]

I keď umelú inteligenciu pokladáme za niečo nové a bytostne sa viažuce na súčasnú modernú – primárne informačnú – spoločnosť, ide o fenomén, ktorý oslovoval mysliteľov a vizionárov už niekoľko storočí dozadu, možno od počiatkov Prvej priemyselnej revolúcie[48], keďže jedným z podstatných prvkov priemyselnej revolúcie bolo nahradenie ľudskej práce strojmi. Činnosť, ktorú dovtedy manuálne vykonávali ľudia, odrazu vykonávali – a efektívnejšie vykonávali – stroje. S konkrétnou paradigmatickou zmenou vtedajšej spoločnosti sa tak začali otvárať nové horizonty aj v pohľade na zariadenia, ktoré môže človek vytvoriť.

Zaujímavým predstaviteľom tohto snaženia môže byť Kempelenov[49] Turek, mechanický stroj zostrojený v roku 1771, ktorý bol vydávaný za šachový automat. Stroj, ktorý bol prezentovaný ako automat, dokázal zohrať dobrú šachovú partiu proti ľudskému protihráčovi. V skutočnosti však išlo o mechanické zariadenie ukrývajúce skutočného šachistu, pričom podvod vyšiel na povrch až v roku 1857, t.j. niekoľko rokov po požiari, pri ktorom stroj zhorel (r. 1854).[50]

Nerozoberajúc mechanické prevedenie Turka[51] je možné si uvedomiť **ideu vytvorenia**

47 doc. Ing. Ľuboš MAGDOLEN, CSc., SjF STU.

48 V rámci Prvej priemyselnej revolúcie sa ako dôsledok pokroku vo výrobných nástrojoch (osobitne v textilných manufaktúrach) a následne vynálezu parného stroja začína industrializácia a masívne využívanie strojov, ktoré nahrádzali ručnú prácu.
Priemyselná revolúcia [on-line]. [cit. 20. augusta 2020].
Dostupné na internete: <https://sk.wikipedia.org/wiki/Priemyseln%C3%A1_revol%C3%BAcia>

49 Barón Wolfgang Kempelen, pochádzajúci z nemecky hovoriacej rodiny, bol uhorský vysoký štátny úradník, polytechnik a vynálezca, ktorý sa narodil 23. januára 1734 v Bratislave.
Wolfgang Kempelen [on-line]. [cit. 20. augusta 2020].
Dostupné na internete: <https://sk.wikipedia.org/wiki/Wolfgang_Kempelen>

50 *Turek (stroj)* [on-line]. [cit. 20. augusta 2020].
Dostupné na internete: <https://sk.wikipedia.org/wiki/Turek_(stroj)>

51 V súčasnosti Amazon ponúka službu *Mechanical Turk* ako „trhovisko pre prácu, ktorá vyžaduje ľudskú inteligenciu“. Mechanical Turk od Amazonu spája záujemcov s úlohami ťažko realizovateľnými počítačmi so zamestnancami, ktorí za poplatok tieto úlohy vykonajú. Takto sa napr. pripravujú veľké množiny dát (sorting and labeling in datasets) pre učenie a testovanie systémov umelej inteligencie.

stroja, ktorý hrá šach a dokáže sa v tejto hre, ktorá bola dlhé roky považovaná za prejav inteligencie, postaviť človekovi.

Tieto tendencie boli následne umocnené s príchodom Druhej a osobitne Tretej priemyselnej revolúcie, v rámci ktorej nové *mysliace stroje* (t.j. počítačové systémy) boli stále viac schopné vykonávať koncepčné, riadiace i administratívne funkcie a koordinovať tok výroby, od ťažby surovín po predaj a distribúciu konečných výrobkov i služieb.[52] Technologické zariadenia a informačné systémy sa tak v podstatnej miere začali podieľať na fungovaní spoločnosti a jej premene na spoločnosť informačnú[53].

Príchod Tretej priemyselnej revolúcie (známej tiež ako digitálnej), ktorý nastal v polovici 20. storočia, znamenal aj úsvit umelej inteligencie a jej systematického bádania.

Digitálna revolúcia, ktorá bola počiatkom informačného veku a premeny spoločnosti na spoločnosť založenú na vedomostiach, sa v kontexte permanentnej technickej revolúcie[54] podieľa na formovaní v súčasnosti nastupujúcej Štvrtej priemyselnej revolúcie (známej ako *Industry 4.0*). Je charakterizovaná zlúčením technológií, ktoré stierajú hranice

---

MITCHELL, *Artificial Intelligence*, s. 85.

52 Druhá priemyselná revolúcia spadá do obdobia druhej polovice 19. storočia až po 1. svetovú vojnu. V tomto čase sa masívnym spôsobom rozrastá priemyselná výroba a prichádza k elektrifikácii spoločnosti.
Tretia priemyselná revolúcia, ktorá sa datuje niekedy od polovice 20. storočia a je tiež známa ako digitálna, je obdobím nástupu digitálnych technológií, počítačov a moderných foriem spracovania informácií i komunikácie. Digitálna revolúcia bola začiatkom informačného veku a premeny spoločnosti na spoločnosť založenú na vedomostiach.
HUMBER, M. *Technology and Workforce: Comparison between the Information Revolution and the Industrial Revolution* [on-line]. Berkeley: University of California, 2007, s. 2-3. [cit. 20. augusta 2020]. Dostupné na internete: <http://infoscience.epfl.ch/record/146804/files/InformationSchool.pdf>

53 ŠANTAVÝ, *Analýza virtuálneho sveta v kontexte náboženskej formácie, pôsobenia a života Cirkvi* [on-line], s. 19-20. [cit. 21. augusta 2020].
Dostupné na internete: <https://peter.santavy.cloud/data/uploads/docs/analyza_virtualneho_sveta_v_kontexte_nabozenskej_formacie_posobenia_a_zivota_cirkvi.pdf>

54 ŠANTAVÝ, *Analýza virtuálneho sveta v kontexte náboženskej formácie, pôsobenia a života Cirkvi* [on-line], s. 23-24. [cit. 20. augusta 2020].
Dostupné na internete: <https://peter.santavy.cloud/data/uploads/docs/analyza_virtualneho_sveta_v_kontexte_nabozenskej_formacie_posobenia_a_zivota_cirkvi.pdf>

medzi fyzickými, digitálnymi a biologickými sférami s dôrazom na celkovú transformáciu spoločnosti na spoločnosť znalostnú a informačnú, pričom dôležitú úlohu zohráva nástup systémov a prvkov umelej inteligencie do takmer všetkých oblastí fungovania spoločnosti.[55]

1.1. Úsvit umelej inteligencie

Ako už bolo spomenuté, úsvit umelej inteligencie nastával ruka v ruke s príchodom digitálnej revolúcie a s rozvojom elektronických výpočtových systémov.

Vtedajší vizionári a inventori v oblasti informačných technológií prekračovali základný i aplikovaný výskum v oblasti matematiky a informatiky pohľadom skôr filozofickým či futurologickým so základnou otázkou, kam až môže vývoj počítačových systémov zájsť.

Keďže informačné a komunikačné technológie (IKT) sa čoraz viac etablovali ako *mysliace stroje* (vysvetlenie a významové zaradenie je spomenuté v predchádzajúcej kapitole), k úvahám o umelej inteligencii bol už len malý krok, ktorý viacerí s nadšením a entuziazmom neváhali vykonať.

Medzi dôležité artefakty tohto obdobia bezpochyby patrí tzv. **Turingov test**, ktorého cieľom bolo zistiť, či je nejaký stroj inteligentný, resp. dokáže myslieť. Turingov test je založený na jednoduchej premise: ak dokáže človek aspoň päť minút konverzovať s nejakým respondentom bez toho, že by zistil, že ide o stroj (nie človeka), tak tento stroj (systém, počítač,...) úspešne prejde testom.[56] Inak povedané, dokáže stroj napodobniť ľudské myšlienky, dokáže myslieť?

---

55 *Prečo Industry 4.0* [on-line]. [cit. 20. augusta 2020].
Dostupné na internete: <http://industry4.sk/o-industry-4-0/co-je-industry-4-0/>

56 V Turingovom teste osoba v úlohe pýtajúceho sa dáva otázky dvom respondentom, s ktorými nie je v priamom kontakte, pričom jedným respondentom je človek a druhým stroj. Komunikácia je výlučne textová, trvá päť minút a rozsah konverzačných tém nie je obmedzený.
Ak pýtajúci sa spolu s hodnotiacou komisiou (hodnotiť odpovede môže viac osôb) nedokáže podľa odpovedí respondentov rozlíšiť, ktorý z nich je človek a ktorý stroj, potom tento stroj možno považovať za inteligentný. Prelomovou hranicou sa považuje 30% úspešnosť, teda minimálne traja z desiatich hodnotiteľov musia byť presvedčení, že komunikácia prebieha s človekom a nie strojom, prípadne nedokážu rozlíšiť človeka od stroja.
Por. *Turingov test* [on-line]. [cit. 29. januára 2021].
Dostupné na internete: <https://sk.wikipedia.org/wiki/Turingov_test>

Test sa prvý krát podarilo zvládnuť v roku 2014 počítačovému programu Eugene, ktorý simuloval konverzáciu trinásť ročného chlapca.[57] A po ňom nasledovali ďalšie úspešné programy, čo len dokazuje, že Turingov test nie je dostatočným kritériom pre umelý systém, ktorý rozmýšľa ako človek. **Základným problémom – a jeho podstata sa týka prakticky takmer všetkých systémov umelej inteligencie dneška – je skutočnosť, že i keď nejaký systém dokáže odpovedať tak, akoby konverzácii rozumel, neznamená to, že je inteligentný a že konverzácii aj skutočne rozumie**.[58] [59] [60]

Každopádne **Alan Turing[61], priekopník a otec modernej počítačovej vedy, ktorý svoj**

57 *First Turing Test success marks milestone in computing history* [on-line]. [cit. 29. januára 2021]. Dostupné na internete: <https://phys.org/news/2014-06-turing-success-milestone-history.html>

58 Realitu tohoto problému vystihuje aj argument čínskej izby, ktorý bol predložený filozofom Johnom Searlom v roku 1980. Ide o myšlienkový experiment názorne vyjadrujúci, že **schopnosť zmysluplne odpovedať na položené otázky nie je dostatočná na preukázanie schopnosti otázkam porozumieť a myslieť**.
SEARLE, R. J. *Minds, Brains, and Programs* In: *The Behavioral and Brain Sciences.* [on-line]. Cambridge University Press, 1980, zv. 3. [cit. 30. januára 2021].
Dostupné na internete: <https://web.archive.org/web/20071210043312/http://members.aol.com/NeoNoetics/MindsBrainsPrograms.html>

59 Nielen program Eugene, ale aj oveľa sofistikovanejšie systémy AI, napr. Watson od IBM, vytvárajú chyby, ktoré sú však často iné, než robia v podobných situáciách ľudia. Sú také „neľudské“, ľudsky nepochopiteľné. Je však zaujímavé, že jedným z faktorov, ktorý dokázal hodnotiteľov Eugene presvedčiť, bola schopnosť robiť chyby, ktoré sú podobné ľudským chybám.

60 V súčasnosti sa pri systémoch spracovania jazyka, inteligentných asistentoch a pod. používa na meranie schopnosti porozumieť komunikácii tzv. SQuAD (The Stanford Question Answering Dataset), ktorý bol v roku 2016 predstavený výskumníkmi Stanfordskej univerzity.
RAJPURKAR, P. et al. *SQuAD: 100 000+ Questions for Machine Comprehension of Text,* In: *Proceeding of the 2016 Conference on Empirical Methods in Natural Language Processing.* 2016, 2382-92.

61 Alan Turing (1912-1954) bol britský matematik, logik a kryptograf, ktorý súbežne (nezávisle od seba) s matematikom Johnom von Neumannom vytvoril princípy dnešných elektronických počítačov. Turing sa už počas II. svetovej vojny preslávil skonštruovaním zariadenia, ktoré umožnilo prekonať nemecké šifrovacie zariadenie Enigma. Povojnovú prácu na konštrukcii digitálnych počítačov rozšíril o štúdium neurónových sietí, uplatnenie matematických metód v biológii a skúmanie vízie mysliacich strojov.
Alan Turing [on-line]. [cit. 30. januára 2021].
Dostupné na internete:
<https://en.wikipedia.org/wiki/Alan_Turing#Pattern_formation_and_mathematical_biology>
The Turing Digital Archive [on-line]. [cit. 30. januára 2021].
Dostupné na internete: <http://www.turingarchive.org/>

test predstavil v roku 1950, ním akoby položil základy umelej inteligencie a odštartoval niečo, čo sa už nedalo zastaviť.

Za skutočný zrod umelej inteligencie sa však považuje **Dartmouthský seminár** – dvojmesačný výskumný projekt zaoberajúci sa umelou inteligenciou, ktorý sa uskutočnil v lete 1956 v Dartmouth college v USA.[62] V podstate išlo o pracovný seminár, ktorý organizoval mladý matematik John McCarthy v spolupráci s Marvinom Minskym (bývalým kolegom zo štúdií, ktorý s ním zdieľal fascináciu inteligentnými počítačmi a neskôr sa stal zakladateľom laboratória umelej inteligencie na MIT), Claudem Shannonom (kolegom z Bell Labs/IBM a vynálezcom teórie informácií) a Nathanielom Rochesterom (pionierom v oblasti elektroniky a architektom viacerých počítačov IBM).

Prvé použitie, či **vynájdenie termínu *umelá inteligencia*** sa pripisuje práve Johnovi McCarthymu, ktorý sa tak chcel vyhraniť voči príbuznej, čerstvo sa rozvíjajúcej oblasti kybernetiky.[63] [64] Je pritom zaujímavé, že použitie prívlastku umelá sa vtedy nepáčilo prakticky nikomu na Dartmouthskom seminári, keďže mali na mysli pravú, skutočnú inteligenciu. Avšak v pracovných debatách ju bolo treba nejako nazvať a tak ju nazvali umelá inteligencia (artificial intelligence).[65] I táto drobnosť vyjadruje počiatočné snahy uchopiť novú oblasť, ktorá sa s rozvojom elektronických počítačov, matematických metód a poznania v oblasti biológie javila ako síce neznáma, no veľmi lákavá.

---

62 *Dartmouth workshop* [on-line]. [cit. 1. februára 2021].
Dostupné na internete: <https://en.wikipedia.org/wiki/Dartmouth_workshop>

63 Kybernetika je vo všeobecnosti veda o skúmaní, riadení, regulácii a komunikácii v dynamických systémoch v technike, biologickej sfére a spoločnosti. Predmetom kybernetiky je spracovanie informácií pre potreby riadenia a jej metódami sú systémový prístup a modelovanie pri riešení problémov. *Kybernetika* [on-line]. [cit. 3. februára 2021].
Dostupné na internete: <https://sk.wikipedia.org/wiki/Kybernetika>

64 Na margo vyhranenia sa voči kybernetike však treba povedať, že umelá inteligencia a kybernetika vo svojom rozvoji majú styčné plochy a spoločné oblasti, napr. robotika, automatizácia, počítačové videnie a pod. Autor tohto textu, absolvujúc štúdium technickej kybernetiky na Elektrotechnickej fakulte SVŠT (aktuálne pôsobiacej pod názvom Fakulta elektrotechniky a informatiky STU), v rámci už vtedajšieho kybernetického odboru navštevoval aj základy neurónových sietí a vo svojej diplomovej práci riešil spracovanie obrazu metódami matematickej morfológie, t.j. metódami, ktoré využíva nielen robotika a automatizované systémy, ale napr. i vstupná vrstva konvolučných sietí používaných na rozpoznávanie obrazu v moderných systémoch umelej inteligencie.

65 NILSSON, N. J., MCCARTHY, J. *A Biographical Memoir*. Washington D.C.: National Academy of Sciences, 2012.

V podkladoch, ktoré boli súčasťou žiadosti o finančnú podporu na realizáciu seminára adresovanej the Rockefeller Foundation, organizátori predstavili návrh, že „**všetky aspekty učenia, resp. akékoľvek iné črty inteligencie môžu byť v princípe tak detailne popísané, že je možné vytvoriť stroje, ktoré by inteligenciu mohli simulovať**“.[66]

Detailné predstavenie tohoto návrhu v sebe zahŕňalo celú množinu tém, ktoré mali byť na seminári diskutované: **strojové spracovanie ľudskej reči, neurónové siete, strojové učenie, abstraktné koncepty a myslenie, kreativita**... Treba povedať, že uvedené témy definujú oblasti skúmania a rozvoja umelej inteligencie dodnes.

Z dnešného uhľa pohľadu – vnímajúc problémy, s ktorými sa pri tvorbe systémov umelej inteligencie potýkajú už celé generácie odborníkov z rozličných oblastí a uvedomujúc si, že najvýkonnejšie počítače v roku 1956 boli milión krát pomalšie, než dnešné bežné chytré mobilné telefóny – je zarážajúci optimizmus, ktorý účastníci seminára ohľadom dosiahnutia umelej inteligencie mali: „myslíme si, že významný pokrok v riešení jedného alebo viacerých z týchto problémov sa dá dosiahnuť, ak starostlivo vybraná skupina vedcov bude na tom spoločne pracovať počas jedného leta“.[67]

Z dnešného uhľa pohľadu vieme, aký nereálny a idealistický pohľad účastníci seminára mali.[68] Nech sa však v nasledovných rokoch a desaťročiach dialo čokoľvek, v roku 1956

66 MCCARTHY, J. et al.: *Proposal for the Dartmouth Summer Research Project in Artificial Intelligence*. In: *AI Magazine.* [on-line]. 1955, 27(4). [cit. 3. februára 2021].
Dostupné na internete: <https://doi.org/10.1609/aimag.v27i4.1904>

67 MCCARTHY, et al.: *Proposal for the Dartmouth Summer Research Project in Artificial Intelligence*. [on-line]. [cit. 3. februára 2021].
Dostupné na internete: <https://doi.org/10.1609/aimag.v27i4.1904>

68 McCarthy na začiatku šesťdesiatych rokov zakladá Standfordský projekt umelej inteligencie „s cieľom vytvorenia plne inteligentného stroja v rámci dekády“.
MORAVEC, H. *Mind Children: The Future of Robot and Human Intelligence*. Cambridge, Mass.: Hardvard University Press, 1988, s. 20.
Budúci laureát Nobelovej ceny v tom istom čase predikuje: „Do dvadsať rokov budú stroje schopné vykonávať akúkoľvek prácu, ktorú môže vykonávať človek“.
SIMON, H. A. *The Shape of Automation for Men and Management*. New York: Harper&Row, 1965, s. 90.
Minského predikcia hovorila podobne nádejne: „v rámci jednej generácie budú problémy spojené s vytvorením umelej inteligencie v podstate vyriešené“.
MINSKY, M. L. *Computation: Finite and Infinite Machines*. Upper Saddle River, N.J.: Prentice Hall, 1967, s. 2.

prakticky započal rozvoj nového vedného odboru, ktorý na Dartmouthskom seminári dostal nielen meno, ale i solídny náčrt základných cieľov.[69]

V kontexte zamerania tohoto textu je zaujímavé primárne inšpirovanie sa realizácie AI biologickými systémami a porovnávanie AI s ľudskou inteligenciou. I keď prvotný pohľad bol čiste biologicko-matematický, moderný rozvoj systémov AI a dosah ich použitia v rôznych oblastiach života priniesol veľké množstvo otázok, ktoré sa dotýkajú *koexistencie* AI a človeka, resp. spoločnosti. Ide teda o otázky psychologické, sociologické a etické. Inak povedané, **ak chceme riešiť pokročilé systémy umelej inteligencie a ich použitie, nie je možné stavať len na matematických a biologických základoch, ale musíme si uvedomiť, že akákoľvek inteligencia – ak sa má takto nazývať – tieto základy bytostne presahuje**.

1.2. Definícia pojmov

Už debata o termíne umelá inteligencia naznačuje otázky, ktoré si od čias Dartmouthského seminára bádatelia v oblasti AI znovu a znovu kladú: Koľko nám toho ešte zostáva zdolať do vytvorenia skutočne inteligentného stroja? Bude vytvorenie plnej umelej inteligencie vyžadovať reverzné inžinierstvo ľudského mozgu v celej jeho komplexite, alebo nájdeme skratku, nejakú super šikovnú množinu dnes ešte neznámych algoritmov, ktoré sa budú podieľať na vytvorení plnej umelej inteligencie? A čo je podstatné – **vieme vôbec definovať inteligenciu[70] a čo to vlastne znamená „plná inteligencia“?**

Voltairova výzva „definujte svoje termíny ... lebo inak nikdy jeden druhému neporozumieme“ je výzvou pre kohokoľvek, kto v oblasti umelej inteligencie tvorí, alebo o nej hovorí.[71] Pretože už základné pojmy ako inteligencia, myslenie, poznanie, vedomie

---

I napriek úžasnému pokroku – osobitne v posledných dekádach - žiadna z týchto predpovedí sa doteraz nenaplnila...

69 Štyria z účastníkov, John McCarthy, Marvin Minsky, Allen Newell a Herbert Simon sa nazývajú „veľká štvorka“ zakladateľov.

70 Novozélandský morálny filozof a psychológ James Flynn definuje inteligenciu ako **súbor zručností, medzi ktoré patrí abstraktné a logické myslenie, ďalej predstavivosť, a teda aj schopnosť uvažovať o hypotetických možnostiach, a tiež aj jazykový cit**.
Por. FLYNN, J. *What Is Intelligence?: Beyond the Flynn Effect*. Cambridge University Press, 2009.

71 Súčasnosť dáva tomu za pravdu, keď sa spojenie umelá inteligencia používa ako buzzword – módny termín – zahŕňajúci všetko možné od inteligentných asistentov a expertných systémov, cez jednoduché

a city sú pre potreby bádania v oblasti umelej inteligencie zle definované. Marvin Minsky preto pre tieto pojmy razí termín „suitcase word“.[72] Každý z týchto pojmov je ako veľký kufor, obsahujúci spleť rôznych významov a možností. Preto i umelá inteligencia participuje na tomto probléme a nadobúda tak rôzne významy podľa konkrétnych kontextov.

Ako ľudia tak nejako podvedome vieme rozlíšiť, čo je a čo nie je inteligentné (napr. človek vs. kameň na ceste,...) a vieme takto aj inteligenciu porovnávať (napr. človek vs. bocian,...). Dokonca máme vytvorené aj merítka pre ľudskú inteligenciu (IQ) a naviac vieme rozlišovať jej rôzne dimenzie: emocionálnu, verbálnu, logickú, sociálnu, atď.

Takže inteligenciu kategorizujeme binárne (niečo je alebo nie je inteligentné), kontinuálne (jeden objekt je inteligentnejší ako druhý) a multidimenzionálne (inteligencia v rôznych oblastiach). Pojem inteligencia teda môžeme v Minského poňatí vnímať ako *na prasknutie naplnený kufor*.

Doterajší vývoj systémov AI však ukazuje, že uvedené rozlišovanie v oblasti AI je vo veľkej miere ignorované a dôraz je skôr kladený na dva rôzne smery vývoja – vedecký a praktický.

Na strane vedeckého vývoja sa výskumníci snažia pochopiť prirodzený, t.j. biologický mechanizmus inteligencie a tento následne implementovať v počítačových systémoch.

Praktický smer vývoja sa snaží skôr jednoduchou a pragmatickou cestou vytvoriť také počítačové systémy, ktoré dokážu vykonávať úlohy rovnako alebo lepšie ako ľudia, pričom pramálo záleží na tom, či tieto programy myslia spôsobom ľudského myslenia.

Správa z roku 2016 stanfordskej *100-ročnej štúdie o umelej inteligencii*[73] **definuje umelú inteligenciu ako oblasť počítačových vied, ktorá študuje vlastnosti inteligencie syntetizovaním inteligencie**.[74]

---

algoritmy a aplikácie umelej inteligencie, autonómne a robotické systémy, až po sofistikované systémy typu AlphaGO a IBM Watson.

72 MINSKY, M. L. *The Emotion Machine: Commonsense Thinking, Artificial Intelligence, and the Future of the Human Mind*. New York: Simon & Schuster, 2006, s. 95.

73 Stanfordská 100-ročná štúdia o umelej inteligencii (AI100) je projekt, ktorý sa začal v decembri 2014 s ambíciou pravidelne vyhodnocovať AI panelom odborníkov raz za 5 rokov. Ide o aktivitu zameranú na pravidelnú analýzu súčasného stavu rozvoja umelej inteligencie a jej vplyvu na ľudí.

74 SIMON, H. A. *Artificial Intelligence: An Empirical Science*. In *Artificial Intelligence* 77. 1995, č. 2, s. 95-

100-ročná štúdia o umelej inteligencii (AI100) v správe z roku 2021 dopĺňa alternatívnu definíciu AI: **umelá inteligencia je o vytváraní systému, ktorý vykazuje také správanie, o ktorom si myslíme, že vyžaduje inteligenciu.**[75]

Aktuálna neexistencia presnej definície môže byť aj výhodou pre akcelerovanie celého odvetvia umelej inteligencie v rôznych smeroch vývoja a aplikovania systémov AI. Neexistencia presnej definície AI a viac menej ignorancia základných kategorizácií sú živnou pôdou pre veľmi rôznorodý a kreatívny rozvoj tejto oblasti, keďže bádateľov vedie len hrubý zmysel pre toto smerovanie a snaha etablovať sa, či dostať sa do tejto oblasti.[76]

1.3. Základné vlastnosti a delenie systémov umelej inteligencie

Na základe doterajšej debaty ohľadom definície a kategorizácie inteligencie – systémy, ktoré by boli schopné vykazovať inteligentné správanie, musia spĺňať dve základné vlastnosti:

- **autonómnosť** – schopnosť samostatne konať, t.j. schopnosť systému vykonávať úlohy v komplexnom prostredí bez neustáleho vedenia používateľom.
- **adaptívnosť** – schopnosť sa prispôsobovať, t.j. schopnosť zlepšovať svoj výkon (a schopnosti) učením sa zo skúseností.

Z predchádzajúcej kapitoly so pripomeňme, že o inteligencii môžeme hovoriť v kategóriách kontinuity (jeden objekt je inteligentnejší ako druhý) a viacerých dimenzií (inteligencia v rôznych oblastiach).

Preto i systémy môžu byť adaptívne a autonómne len v určitej oblasti, t.j. sú schopné riešiť určité úlohy „inteligentným spôsobom“, pričom v ostatných oblastiach zlyhávajú. Takéto systémy nazývame **úzko špecializované umelé inteligencie (narrow AI)**[77]. Ide teda

---

127.
One Hundred Year Study on Artificial Intelligence. [on-line]. AI100, 2016, s. 13. [cit. 14. novembra 2021]. Dostupné na internete: <https://ai100.stanford.edu/2016-report>

75 *One Hundred Year Study on Artificial Intelligence*. [on-line]. AI100, 2021, s. 78. [cit. 15. novembra 2021]. Dostupné na internete: <https://ai100.stanford.edu/2021-report>

76 MITCHELL, *Artificial Intelligence*, s. 20.
One Hundred Year Study on Artificial Intelligence. [on-line]. AI100, 2016, s. 12. [cit. 14. novembra 2021]. Dostupné na internete: <https://ai100.stanford.edu/2016-report>

77 Používa sa skratka **ANI – Artificial Narrow Intelligence.**

o vysoko špecializované systémy, ktoré sú optimalizované na zvládnutie konkrétnej úlohy, resp. množiny úloh.[78]

Všeobecná umelá inteligencia (general AI)[79] je systém, ktorý dokáže zvládnuť akúkoľvek intelektuálnu úlohu. V podstate ide o umelú inteligenciu, ktorá je na úrovni človeka.[80]

Všetky metódy a systémy umelej inteligencie, ktoré dnes používame, spadajú do kategórie špecializovaných systémov AI (narrow AI), pričom súčasný rozvoj v tejto oblasti napreduje míľovými krokmi.

Všeobecná umelá inteligencia – i napriek bombastickým titulkom v novinách a niektorým futurologickým predpovediam – patrí v súčasnosti do oblasti sci-fi.

Analogicky rozlišujeme umelú inteligenciu ako *silnú* a *slabú* na základe rozlišovania medzi inteligentnými systémami a systémami, ktoré inteligentne konajú.[81]

Silná umelá inteligencia (strong AI) je skutočne inteligentná a „uvedomelá“, t.j. skutočne aj rozumie tomu, či rieši a vykonáva. Uvedomelá AI je súčasne aj všeobecnou, pretože má schopnosť generalizovať, t.j. zovšeobecňovať a prenášať, či adaptovať naučené schopnosti na iné úlohy (čo mimochodom patrí k základom ľudského myslenia).

Slabá umelá inteligencia (weak AI) vykazuje inteligentné správanie na základe modelov a aplikovaných metód i dát, na ktorých sa učí (je natrénovaná). Ide teda o systémy, ktoré sú zamerané na riešenie konkrétnych úloh a sú závislé na ľudskom vstupe a konfigurácii.[82]

---

78 IBM Deep Blue napr. dokáže poraziť najlepších šachistov sveta, no ak by tento systém mal byť použitý v autonómnych vozidlách, nedokázal by sa ani rozbehnúť na rovnej ceste.

79 Používa sa skratka **AGI – Artificial General Intelligence**.

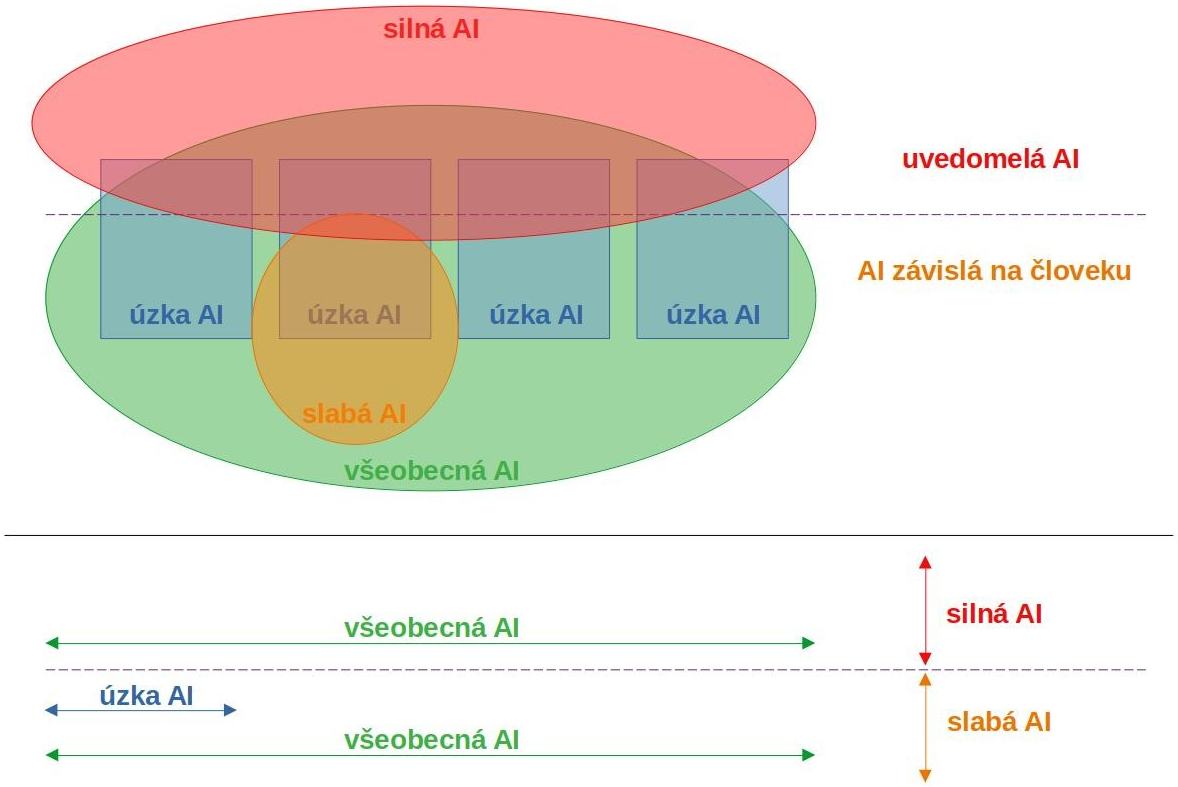
80 Niekedy sa z okruhu AGI samostatne vyčleňuje ešte **ASI – Artificial Super Intelligence**, t.j. umelá inteligencia, ktorá by bola naprieč všetkými oblasťami inteligentnejšia ako človek. Úvahy o ASI sa mnohokrát spájajú s konceptom *singularity v oblasti AI*, v ktorej umelá inteligencia so schopnosťou učiť sa, zlepšovať sa a samostatne sa vyvíjať rýchlo dosiahne a následne extrémne prevýši inteligenciu človeka. Viac o vízii Raya Kurzweila a o singularite je uvedené v poznámke č. 16.

81 Tento rozdiel akcentuje zvýraznený odstavec a poznámky o základnom probléme v kapitole 1.1.: „Základným problémom – a jeho podstata sa týka prakticky takmer všetkých systémov umelej inteligencie dneška – je skutočnosť, že i keď nejaký systém dokáže odpovedať tak, akoby konverzácii rozumel, neznamená to, že je inteligentný a že konverzácii aj skutočne rozumie.“

82 Vytvorenie skutočne dobre fungujúceho systému AI vyžaduje skutočných odborníkov, schopných aplikovať správne metódy i správne nakonfigurovať a parametrizovať daný systém. Pre neznalých to

Slabá AI reprezentuje systémy, ktoré aktuálne máme, t.j. počítačové systémy, ktoré vykazujú inteligentné správanie.

Nasledujúci obrázok vyjadruje dve znázornenia vzájomného vzťahu a pomeru medzi silnou a slabou, všeobecnou a úzko špecializovanou umelou inteligenciou.



Obr. 1. Vzťah medzi silnou a slabou, všeobecnou a úzko špecializovanou umelou inteligenciou.

1.4. Symbolické a subsymbolické systémy ako súčasť anarchie metód

Ak sme trošku medzi riadkami čítali v kapitole 1.1. o úsvite umelej inteligencie a osobitne o Dartmouthskom seminári, môžeme tušiť, že už od počiatkov výskumu existoval nespočet pohľadov na to, akým spôsobom treba uchopiť problematiku umelej inteligencie a akými cestami by sa jej výskum a vývoj mal uberať. Ani nasledovné desaťročia na rozsahu tejto spletitosti neubrali. A preto, vzhľadom na našu nedostatočnosť chápania inteligencie ako

môže vyzerať ako mágia, či voodoo:-)

Por. MITCHELL, *Artificial Intelligence*, s. 98.

takej[83] a z nej prameniacej neschopnosti uchopiť vytvorenie všeobecnej umelej inteligencie, miesto jasnej cesty vývoja AI sa vývoj a následný pokrok dosahuje v takej rozmanitosti a spletitosti metód AI, že môžeme hovoriť o anarchii metód.[84]

Výskum umelej inteligencie tak obsahuje širokú množinu prístupov, ktorá – na prvý pohľad – môže byť pre vývoj systémov AI brzdou. Máme však za to, že v kontexte našej nedostatočnosti chápania inteligencie ako takej, je tento prístup asi jediný možný. Na jednej strane sú tak preskúmané smery a metódy, ktoré k cieľu nevedú, na strane druhej súčasné úspechy v oblasti rozvoja AI stavajú na schopnosti kombinovať rôzne metódy a prístupy,[85] prípadne využiť staršie, javiace sa ako neúspešné, avšak zdokonalené natoľko, že prinášajú kýžené ovocie.[86]

I napriek uvedenej anarchii však môžeme aspoň principiálne – filozoficky rozdeliť túto širokú množinu metód do dvoch za nimi stojacich prístupov:

- symbolická AI
- subsymbolická AI

Symbolická umelá inteligencia ide cestou vytvárania umelej inteligencie na báze ľudského myslenia, t.j. pojmov, slov, fráz (= symboly) a vzťahov medzi nimi. Symbolické systémy na základe definovaných pravidiel a postupov („ak niečo, tak potom toto“) môžu jednotlivé symboly spracovávať a vykonávať priradené úlohy.[87] Pre symbolickú AI sa význam jednotlivých symbolov odvíja od spôsobu ich kombinácie, vzájomných vzťahov

83 Pojednávali sme o tom v kapitole 1.2.

84 LEHMAN, J., CLUNE, J., RISI, S.: *An Anarchy of Methods: Current Trends.* In: *How Intelligence Is Abstracted in AI*. IEEE Intelligent Systems. 2014, 29, č. 6, s. 56-62.

85 Napr. realizácia konvolučných sietí.

86 Jeden konkrétny príklad z praxe: „The return of patch-based self-supervision! It never worked well and you had to bend over backwards with ResNets (I tried). Now with ViT, very simple patch-based self-supervised pre-training rocks! First BeIT, now Masked AutoEncoders i1k=87.8%“ [on-line]. Twitter. [cit. 18. novembra 2021].
Dostupné na internete: <https://twitter.com/giffmana/status/1459092079020285976>
HE, K., CHEN, X., XIE, S., LI, Y., DOLLÁR, P., GIRSHICK, R.: *Masked Autoencoders Are Scalable Vision Learners* [on-line]. Facebook AI Research (FAIR). [cit. 18. novembra 2021].
Dostupné na internete: <https://arxiv.org/abs/2111.06377>

87 MITCHELL, *Artificial Intelligence*, s. 21.

a operácií, ktoré môžu byť nad nimi vykonávané.[88]

Keďže logika je nutnou podmienkou nášho chápania všeobecnej inteligencie, symbolické systémy sa snažia logickými operáciami riešiť rôznorodé úlohy vyžadujúce nejaký stupeň inteligencie. Striktná formálnosť logického uvažovania umožňuje algoritmizovať ho a previesť do strojovej formy.

Počas prvých dvoch či troch dekád výskumu umelej inteligencie symbolické systémy prevládali. Jednou z prvých implementácií symbolickej AI bol General Problem Solver (GPS)[89], ktorý vedel riešiť také problémy, ako napr. hlavolam Misionári a kanibali[90]. Jedným z vrcholov symbolických systémov sú expertné systémy,[91] ktoré využívali (a dodnes využívajú) človekom vytvorené pravidlá pre riešenie úloh spojených s lekárskou diagnostikou či právnymi rozhodnutiami, alebo zadania v spravodajských službách[92] atď.

Prvé symbolické systémy využívali tzv. výrokovú logiku, no jej aplikácia v systémoch AI rýchlo narazila na limity a obmedzenia pri snahe logicky popísať všetky možnosti, ktoré pri vykonávaní logických operácií nad symbolmi reálneho sveta môžu nastať. Jednoducho povedané, výrokovou logikou nedokážeme popísať tieto úlohy v dostatočne všeobecnej rovine a algoritmy umelej inteligencie s výrokovou logikou nedokážu zovšeobecňovať bez predchádzajúceho masívneho natrénovania, resp. popisu všetkých možných situácií.[93]

Oveľa sľubnejšie sa javia systémy využívajúce logiku prvého rádu, ktoré ponúkajú formálny jazyk schopný širších a všeobecnejších formulácií pre formulovanie logických operácií a tým aj rôznych aspektov inteligencie. **Zatiaľ čo výroková logika stavia**

---

88 MITCHELL, *Artificial Intelligence*, s. 23.

89 NEWELL, A., SIMON, H. A. *GPS: A Program That Simulates Human Thought*. Santa Monica, California: P-2257, Rand Corporation, 1961.

90 *Missionaries and cannibals problem* [on-line]. [cit. 18. novembra 2021]. Dostupné na internete: <https://en.wikipedia.org/wiki/Missionaries_and_cannibals_problem>

91 I keď pre autora už v tých časoch boli neurónové siete a subsymbolická AI oveľa príťažlivejšia, na konci deväťdesiatych rokov bol súčasťou tímov, ktoré pre reálne použitie zvažovali nasadenie expertných systémov ako v tom čase jasnej a reálnej voľby.

92 Využívajú sa napr. systémy na inteligentné vyhľadávanie a hlbokú analýzu dát (data mining) alebo na rozpoznávanie objektov (supervector machine).

93 Analogický problém majú aj na štatistike a pravdepodobnosti postavené metódy s porovnateľnými vyjadrovacími schopnosťami.
Por. RUSSELL, *Human Compatible*, s. 270.

na výrokoch, ktoré môžu byť pravdivé alebo nepravdivé, logika prvého rádu stavia na rôznorodých vzťahoch medzi objektami. V rámci riešenia úloh, ktoré boli založené na presných a istých informáciách, systémy využívajúce logiku prvého rádu dosiahli veľkú dokonalosť.

Pri snahe riešiť zložitejšie úlohy, ktoré mali napr. nepresné vstupy, či vyžadovali rozpoznávanie objektov v reálnych obrazových scénach s vizuálnym šumom a pod., však symbolické systémy zlyhávali. **Nádejné AI systémy postavené na logike prvého rádu tak pohoreli na neschopnosti zvládnuť neisté a nejednoznačné informácie**.[94]

Popritom treba uviesť, že symbolické systémy umelej inteligencie stále existujú, v niektorých oblastiach AI sa stále využívajú a ovocie ich doterajšieho vývoja, resp. kombinácia s rôznymi algoritmami subsymbolických metód je zvažovaná ako reálny vklad do diskusie o vývoji nových systémov ANI a AGI.[95]

Subsymbolická umelá inteligencia a jej vznik boli inšpirované pokrokom v neurovede. Subsymbolický prístup k AI sa snaží uchopiť naše myšlienkové procesy, ktoré by sme mohli nazvať niekedy nevedomými, či automatickými, a ktoré sú základom tzv. rýchleho vnímania (fast perception), čo využívame napr. pri rozpoznávaní tvárí alebo identifikácii hovorených slov.

Subsymbolické programy umelej inteligencie tak neobsahujú súbor presných softvérových postupov na úrovni logického myslenia, ale sú tvorené len stohom rovníc, pre nezainteresovaného len húštinou ťažko interpretovateľných operácií s číslami. **Subsymbolické systému sú navrhnuté tak, aby sa učili vykonávať úlohy na základe dát.[96]**

Symbolická AI – veľmi zjednodušene povedané – sa pomocou matematickej logiky snaží emulovať procesy myslenia[97], subsymbolická AI – znovu zjednodušujúc – sa snaží emulovať činnosť mozgu na úrovni neurónov.

Symbolický prístup – využívajúci prísnu logiku a stavajúci na jasne definovaných charakteristikách prostredia a vzťahov medzi objektami činnosti systémov AI – dosahoval

94 Por. RUSSELL, *Human Compatible*, s. 271.

95 Por. MITCHELL, *Artificial Intelligence*, s. 24.

96 Por. MITCHELL, *Artificial Intelligence*, s. 24.

97 Procesy, ktoré u ľudí zahŕňajú zmyslové vnímanie, abstrahovanie pojmov, vytváranie súdov a úsudkov.

svoj vrchol v osemdesiatych rokoch minulého storočia.[98] Tento prístup však dokázal byť úspešný len pri riešení špecifických typov problémov založených na konkrétnych charakteristikách prostredia, vybraných interakciách systémov AI s týmto prostredím a zadanej konkrétnej množine cieľov. Všetko ostatné, čo by mohlo a malo byť predmetom činnosti akejkoľvek inteligencie, však bolo pre tieto systémy tabu – symbolické systémy zlyhávali a zlyhávajú pri riešení problémov, ktoré sa nedajú exaktne popísať a v reálnych prostrediach, ktoré nie je možné deterministicky uchopiť a sú plné *nejasných* informácií.

Preto väčšina moderných implementácií systémov umelej inteligencie, medzi ktoré patria systémy strojového učenia a neurónové siete, vychádza zo subsymbolického prístupu, ktorý sa snaží tieto problémy uchopiť a v určitej miere aj úspešne riešiť.

1.5. Systémy inšpirované činnosťou mozgu na úrovni neurónov

Ako sme v predchádzajúcej kapitole uviedli, **väčšina moderných implementácií systémov umelej inteligencie vychádza zo subsymbolického prístupu, ktorý je inšpirovaný činnosťou mozgu na úrovni neurónov**.

Tento prístup si môžeme objasniť na jednom z prvých príkladov subsymbolického, mozgom inšpirovaného programu AI, ktorý bol na sklonku 50-tych rokov minulého storočia pod názvom **perceptron** vyvinutý psychológom Frankom Rosenblattom.[99]

I keď by sa mohlo zdať, že v dnešnej dobe ide z pohľadu počítačovej vedy o príklad z praveku, perceptron bol nielen míľnikom v oblasti umelej inteligencie, ale predovšetkým principiálne ovplyvnil vývoj subsymbolických systémov a bol starým otcom moderných a úspešných súčasných systémov AI, ktoré poznáme ako hlboké neurónové siete.

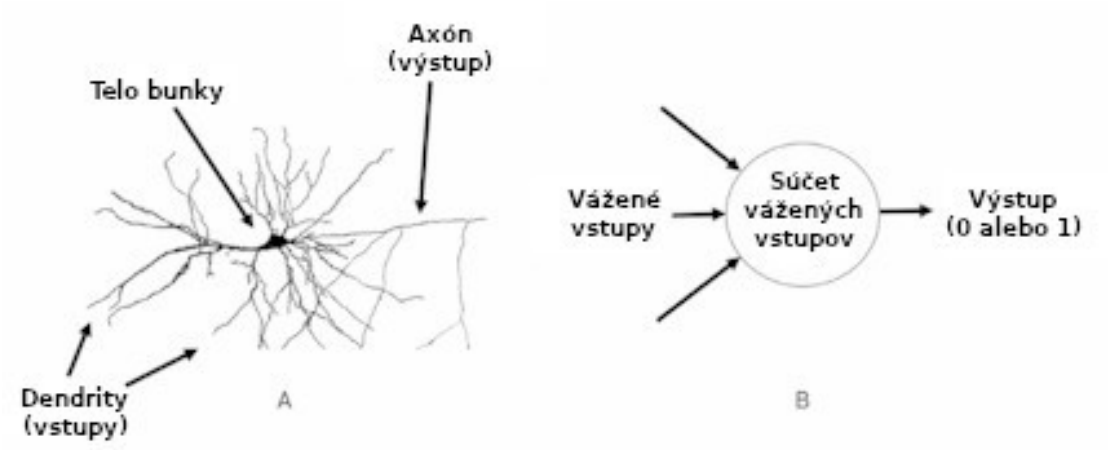
Fungovanie perceptronu bolo principiálne inšpirované spôsobom, ako neuróny spracúvajú informácie: neurón ako mozgová bunka spracúva elektrické, resp. chemické vstupy (vzruchy) z iných neurónov, s ktorými je prepojená. Veľmi zjednodušene povedané, neurón zráta všetky vstupy z ostatných neurónov a ak celková hodnota presiahne určitú hraničnú

98 Išlo o také systémy ako Fifth Generation project japonskej vlády, projekty vládnej agentúry DARPA v USA, prípadne britská Strategic Computing Initiative, ktoré využívali tzv. first-order logic pretavenú do symbolického programovania v programovacom jazyku Prolog.
Por. RUSSELL, *Human Compatible*, s. 271.

99 ROSENBLATT, F. *The Perceptron: A Probabilistic Model of Information Storage and Organization in the Brain*. In: *Psychological Review.* 1958, 65, č. 6, s. 386-408.

hodnotu, neurón vyšle impulz. Pritom je dôležité, že jednotlivé spojenia od ostatných neurónov (synapsy) sú rozlične silné. Keď následne neurón ráta výsledný sumár, vstupy zo silnejších synáps majú väčšiu váhu než vstupy zo synáps slabších. Podľa neurovedcov práve úprava sily jednotlivých spojení medzi neurónmi je jedným z kľúčových aspektov mechanizmu učenia sa mozgu.

Perceptron je v podstate počítačovou simuláciou vyššie uvedeného procesu spracovávania informácií v neurónoch. Túto simuláciu vyjadruje obrázok č. 2., pričom časť A zobrazuje neurón s označenými rozvetvenými vláknami, ktoré privádzajú vstupy do bunky (dendrity) a výstupným kanálom (axon). Časť B je schematickým náčrtom jednoduchého perceptronu, ktorý sčítava vstupy a v prípade, že výsledný súčet je rovný alebo väčší ako nastavená prahová hodnota, výstup je 1 (analogické vyslaniu impulzu u neurónu), v opačnom prípade je hodnota výstupu 0.



Obr. 2: A – neurón v mozgu, B – jednoduchý perceptron.[100]

V schematickom náčrte perceptronu ako simulácie neurónovej bunky je slovným spojením „vážené vstupy“ vyjadrený podstatný aspekt, bez ktorého nie je možné simulovať mechanizmus učenia sa mozgu – a to vyjadrenie sily jednotlivých vstupných dendritov a ich úprava ako súčasť tohoto mechanizmu učenia sa, t.j. vyjadrenie váhy jednotlivých vstupov perceptronu a proces ich modifikácie ako súčasť učenia sa (adaptácie) systému umelej inteligencie.

Možno si ani neuvedomujeme, ako tento jednoduchý koncept využívame pri niektorých svojich rozhodnutiach. Čo nás napríklad môže viesť k zhliadnutiu konkrétneho filmu

100MITCHELL, *Artificial Intelligence*, s. 25, upravené autorom.

v kine? Referencie od priateľov, ktorí tento film už videli, pričom ich názor určite nedávame na rovnakú úroveň – úsudku a vkusu konkrétneho priateľa dôverujeme viac, inému menej. Vytvára sa nám tak množina vážených vstupov, pri ktorých ak pozitívne naladenie pre daný film presiahne určitú úroveň, rozhodneme sa pre jeho zhliadnutie v kine, či nákup v on-line službe. Samozrejme, prichádzajú do úvahy aj ďalšie vplyvy (tzv. hyperparametre), ako napr. reklama na film (*doplnková* konštanta), konkrétne preferencie pre filmový žáner a hercov (variabilita prahovej konštanty, t.j. úrovne, ktorá ak je presiahnutá, rozhodneme sa pre zhliadnutie či nákup filmu) a pod.[101] A pri výbere ďalšieho filmu na zhliadnutie úplne podvedome prichádza k zmenám v dôvere k úsudku jednotlivých našich priateľov – na základe toho, do akej miery bola ich rada pri predchádzajúcom filme relevantná a korešpondujúca s dojmom, ktorý v nás zhliadnutý film zanechal (automatický mechanizmus na zmenu váhy jednotlivých „vstupov“).

Analogicky k uvedenému príkladu bol jednoduchý koncept perceptronu postupne rozvinutý do podoby, v ktorej sú v systémoch AI implementované samoučiace sa mechanizmy na zmenu váhy vstupov, aplikované doplnkové konštanty k súčtu vážených vstupov a variability prahovej hodnoty, atď.

Na rozdiel od systémov symbolickej AI, ktorej procesy „inteligencie“ boli výsledkom presných pravidiel a vzťahov medzi symbolmi (pojmy, frázy, slová,...), **u subsymbolických systémov je všetka ich „inteligencia“ zakódovaná v číslach, ktoré reprezentujú váhy a prahové hodnoty**.[102] Ide teda o stoh rovníc (viď predchádzajúca kapitola), ktoré fungujú na základe správneho nastavenia váh a prahových hodnôt. Tento proces môžeme nazvať učením, pričom systémy sa učia na vzorkách učebných (trénovacích) dát.

Realizácia systémov umelej inteligencie na základe automatického (samostatného) učenia sa z existujúcich vzoriek, dát, prípadne skúseností[103] sa nazýva strojovým

---

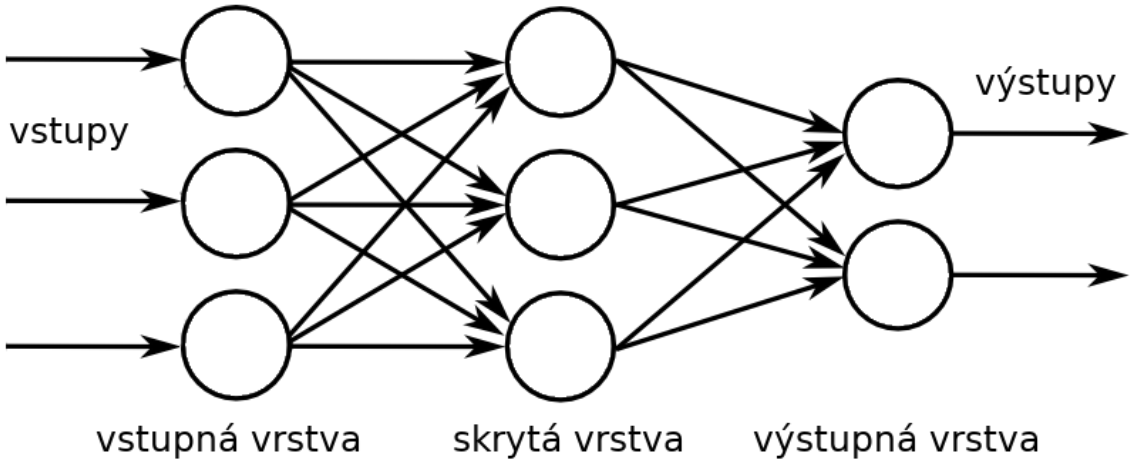
101 Inšpiroval som sa príkladom prof. Melanie Mitchell, ktorý som rozviedol o ďalšie aspekty systémov AI. Por. MITCHELL, *Artificial Intelligence*, s. 25.

102 Základom pre moderné neurónové siete boli tzv. konekcionistické siete (usporiadanie a prepojenie viacerých „neurónov“), v ktorých termín *connectionist* vyjadruje ideu, že „poznanie“ v týchto sieťach sa nachádza vo vážených spojeniach medzi jednotlivými uzlami.
RUMELHART, D. E., MCCLELLAND, J. L. *Parallel Distributed Processing.* Bradford Book, 1986, zv. 1/2.

103 Pod skúsenosťou môžeme rozumieť napr. dáta zo senzorov robotického systému, ktorý práve narazil na prekážku a pod.

učením (machine learning). Výsledkom je nachádzanie vzorov vo vstupných dátach a také nastavenie váh a prahových hodnôt, ktoré systému umožní poskytovať relevantné výsledky a rozhodnutia, to znamená adekvátne reagovať na rôzne vstupné hodnoty bez toho, aby bol na ne explicitne naprogramovaný, iba na základe informácií, ktoré sa už naučil.[104] [105]

Analogicky k perceptronu, **simulácia viacerých poprepájaných neurónov tvorí neurónovú sieť**, pričom jednotlivé simulácie neurónov – uzly (node/unit) sú radené do vrstiev (layer). Vstupné dáta sú postupne spracúvané jednotlivými – skrytými vrstvami (hidden layers), ktoré obsahujú skryté uzly. Skryté vrstvy môže predchádzať ešte samostatná vstupná vrstva (input layer). Posledná vrstva, ktorej výstupom je výsledok činnosti neurónovej siete, sa nazýva výstupná (output layer). Uzly medzi jednotlivými vrstvami sú navzájom poprepájané váženými spojeniami.



Obr. 3: Viacvrstvová neurónová sieť.[106]

Ak má neurónová sieť viac než jednu vrstvu, nazýva sa viacvrstvová (multilayered network). Jej príklad je uvedený na obr. č. 3.

Sieť, ktorá obsahuje viac ako jednu skrytú vrstvu, sa nazýva **hlbokou neurónovou sieťou** (deep neural network, prípadne len deep network), ako je uvedené na obr. č. 4.

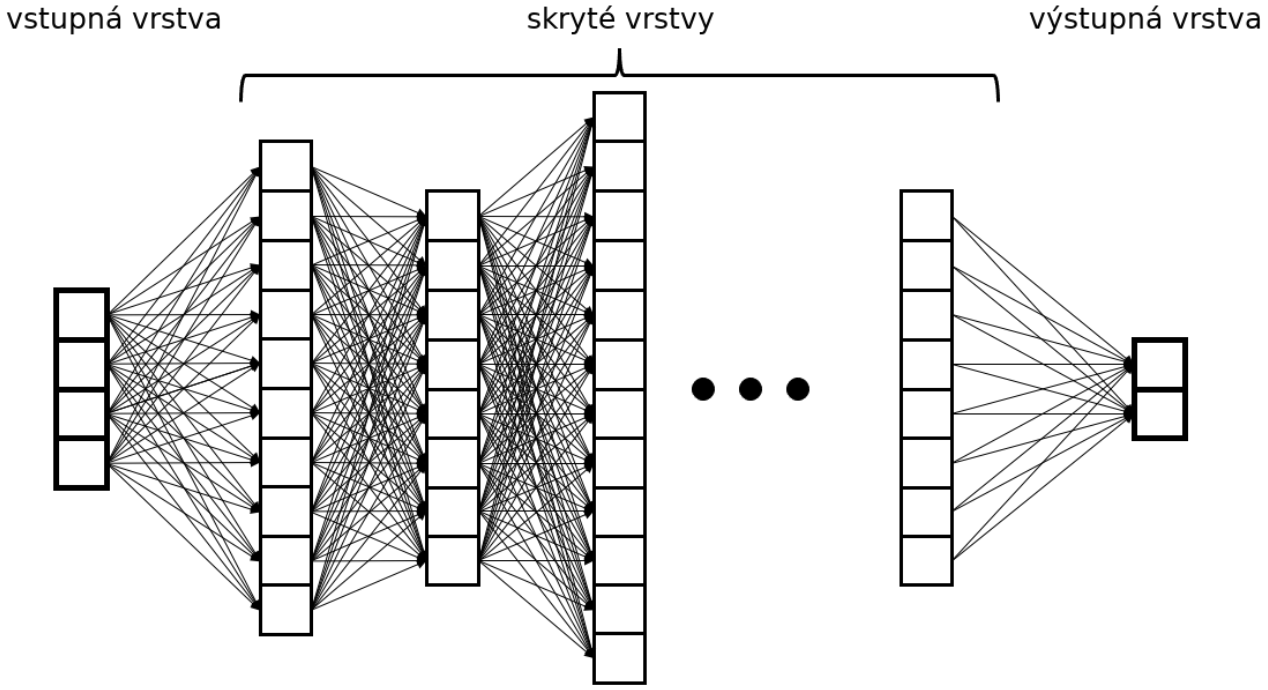
Súčasné sofistikované a najvýkonnejšie systémy umelej inteligencie využívajú typ strojového učenia, ktorý sa nazýva **hlboké učenie** (deep learning). Algoritmy hlbokého

104 Napr. správna detekcia písaného textu, označenie prekážky na ceste, identifikácia hovoreného slova,...

105 Por. *Algoritmy strojového učenia I.* [on-line]. [cit. 3. januára 2022]. Dostupné na internete: <https://umelainteligencia.sk/algoritmy-strojoveho-ucenia/>

106 Wikimedia.org, licencia CC, upravené autorom.

učenia sa učia trénovaním na obrovskom množstve dát prostredníctvom skrytých vrstiev prepojených uzlov, ktoré – ako sme uviedli – sa označujú ako hlboké neurónové siete.



Obr. 4: Hlboká neurónová sieť.[107]

1.6. Základné algoritmy strojového učenia

Podľa toho, ako konkrétne proces učenia prebieha, môžeme algoritmy systémov subsymbolickej AI, resp. strojového učenia rozdeliť do nasledovných skupín:

- učenie s učiteľom (supervised machine learning)
- učenie bez učiteľa (unsupervised machine learning)
- učenie formou odmeňovania (reinforcement learning)

1.6.1. Učenie s učiteľom (supervised machine learning)

Systémy AI využívajúce učenie s učiteľom tvoria v súčasnosti väčšinu systémov strojového učenia. Základom je dostatočne veľká množina správne označených/klasifikovaných trénovacích dát (labeled training dataset) s pozitívnymi a negatívnymi príkladmi. Napr. ak chceme naučiť systém rozpoznávať rôznym spôsobom napísanú číslicu 5, musíme mať veľkú kolekciu napísaných 5-iek, pri ktorých je uvedené označenie (label), že ide o číslicu 5, a tiež veľkú kolekciu iných písaných číslic s označením, že to nie je číslica 5. Tieto

107 Wikimedia.org, licencia CC, upravené autorom.

kolekcie tvoria tzv. trénovacie dáta (training set/dataset). Systém v procese učenia vyhodnocuje každú jednu predloženú číslicu, pričom výsledok sa porovná s jej označením (je – nie je to 5). V prípade nezhody, sa generuje signál (supervision signal), označujúci nesprávny výsledok, resp. mieru nezhody. Na základe takýchto signálov si systém mení nastavenia váh a prahových hodnôt, až kým sa nedostane do stavu, že všetky vstupné dáta z tréningovej množiny nie sú správne vyhodnotené (v našom prípade všetky číslice 5 vyhodnotené ako 5 a ostatné vyhodnotené ako niečo iné).[108]

Ide samozrejme o schematický opis fungovania množiny algoritmov pre supervised machine learning, ktorých presný popis nie je cieľom tohto diela.

1.6.2. Učenie bez učiteľa (unsupervised machine learning)

Učenie bez učiteľa sa používa pri dátach, ktoré neboli vopred klasifikované. Chýba nám teda označenie (label) – správna odpoveď, ktorá klasifikuje vstupné dáta (napr. ide o číslicu 5, na vstupe je trojuholník,...).[109]

Systémy s učením bez učiteľa sa využívajú v prípade, že na vstupe nemáme klasifikované dáta, alebo ide o dáta, ktoré nepoznáme, takže nevieme natrénovať systém na základe dát, ktoré by sme poznali a vedeli ich klasifikovať, resp. označiť. Základné algoritmy využívajúce učenie bez učiteľa teda nie sú schopné určovať správny výstup – ich cieľom je skôr skúmanie a analýza vstupných dát v snahe objaviť a popísať vzory a štruktúry v týchto dátach, teda dozvedieť sa o dátach, resp. z týchto dát niečo viac.

Učenie bez učiteľa sa využíva na riešenie úloh zhlukovania a asociovania.

Cieľom zhlukovania je spájanie dát do skupín, ktoré majú niečo spoločné, napr. segmentácia zákazníkov do skupín s podobnými preferenciami, niektoré problémy spracovania obrazu a detekcie objektov a pod.

---

108 Cyklus prijatia informácie na vstupe – jej spracovania – vyhodnotenia výsledku – následná zmena váhových stavov sa pri zložitejších neurónových sieťach nazýva epochou trénovania systému AI. Pri trénovaní systém prechádza mnohými epochami, ktorých ovocím sú meniace sa váhové stavy i parametre systému a zlepšujúca sa klasifikácia, t.j. lepšie výsledky. Nakoniec systém „konverguje“, t.j. medzi jednotlivými epochami sa váhové stavy takmer vôbec nemenia a neurónová sieť je v princípe v rozsahu trénovacích dát vytrénovaná (naučená).
Por. MITCHELL, *Artificial Intelligence*, s. 80.

109 Por. *Algoritmy strojového učenia II.* [on-line]. [cit. 3. januára 2022]. Dostupné na internete: <https://umelainteligencia.sk/algoritmy-strojoveho-ucenia-ii-ucenie-bez-ucitela/>

Asociovanie sa využíva pri vyhľadávaní asociačných pravidiel, ktoré popisujú množiny dát a vyjadrujú vzťahy medzi nimi. Napr. predajné systémy, ktoré po natrénovaní vedia zákazníkom kupujúcim si určitý produkt ponúknuť relevantné ďalšie produkty (napr. ponuka periférií pri kúpe notebooku), riešenie niektorých problémov pri jazykových prekladoch a pod.

1.6.3. Učenie formou odmeňovania (reinforcement learning)

Ide o osobitnú kategóriu algoritmov strojového učenia, v ktorom sa trénovanie modelu (agenta) realizuje prostredníctvom interakcie s prostredím metódou pokus-omyl. Základom učenia formou odmeňovania (nazývaného aj učenie s posilnením) sú pravidlá, podľa ktorých sa agent môže v danom prostredí správať a odmeňovacia funkcia (funkcia užitočnosti), prostredníctvom ktorej agent vie vyhodnotiť, či vykonané rozhodnutie bolo pre neho správne alebo nie.[110]

Učenie agenta teda neprebieha na žiadnych označených, či neoznačených trénovacích dátach, systém v danom prostredí skúša jednotlivé možnosti a učí sa, ktorá možnosť, resp. kombinácia možností je správna a ktorá nie.

Samozrejme ide len o vyjadrenie princípu – jeho realizácie vo forme algoritmov, ako sú napr. Q-learning či Deep Q-learning sú oveľa sofistikovanejšie a využívajú ich napr. šachové systémy, systémy pre pohyb robotických súprav v prostredí a pod.

Pre správne vytrénovanie systému učením formou odmeňovania treba zvyčajne extrémne veľa iterácií (napr. milióny pokusov).

Toľko aspoň v krátkosti k jednotlivým skupinám algoritmov strojového učenia (detailnejšie rozdelenie je uvedené v prílohách), na ktoré sa teraz môžeme pozrieť trochu z nadhľadu...

1.7. Stačí súčasné strojové učenie, alebo hľadáme ďalej?

Prakticky prvým z algoritmov strojového učenia a ich podmnožiny učenia sa s učiteľom bol Rosenblattov percepton-learning algorithm. Rosenblatt tiež matematicky dokázal, že pre veľmi špecifickú a limitovanú množinu úloh dokáže správne natrénovaný perceptron (resp. podobný systém učenia sa s učiteľom) vykonávať úlohy bez chýb. Avšak pri iných,

110 Por. *Algoritmy strojového učenia III.* [on-line]. [cit. 4. januára 2022]. Dostupné na internete: <https://umelainteligencia.sk/algoritmy-strojoveho-ucenia-iii-ucenie-formou-odmenovania/>

resp. všeobecnejších úlohách umelej inteligencie je úspešnosť takéhoto systému diskutabilná.[111] [112] [113]

Už pri tak jednoduchej realizácii algoritmu strojového učenia je možné si uvedomiť viaceré riziká spojené so systémami umelej inteligencie, napríklad:[114]

- nesprávna zmena váh a prahových hodnôt v procese učenia (napr. priveľké zmeny váhových hodnôt medzi jednotlivými krokmi učenia) môže viesť k nesprávnemu natrénovaniu systému (systém bude dávať nesprávne výsledky v prevádzke).
- u niektorých skupín učenia je pre realizáciu konkrétnych modelov treba vykonať nesmierne veľa iterácií v procese učenia. Obmedzenie možností učenia vedie k modelom, ktoré dávajú nesprávne, resp. v určitých situáciách nesprávne výsledky.
- obmedzenie na množinu úloh, ktoré dokáže takýto systém riešiť.
- praktická nemožnosť previesť naučené hodnoty váh a prahových hodnôt do ľudsky pochopiteľných pravidiel, systém sa tak stáva „čiernou skrinkou“.

Čim zložitejší systém AI (s miliónmi váhových stavov a prahových hodnôt, s viacerými vrstvami neurónových sietí, atď.), tým sú uvedené problémy vo väčšine scenárov vypuklejšie. Vyvstáva tak dôležitá otázka – **sme potom schopní overovať a nastavovať etické pravidlá, resp. všeobecne obmedzenia a mantinely systémov umelej inteligencie?[115]**

---

111 Por. MITCHELL, *Artificial Intelligence*, s. 30.

112 MINSKY, M. L., PAPERT, S. L. *Perceptrons: An Introduction to Computational Geometry*. Cambridge, Mass: MIT Press, 1969.

113 Zložitejšie úlohy samozrejme vyžadujú použitie viacvrstvových neurónových sietí, pre ktoré však v tom čase neexistoval učiaci sa algoritmus, t.j. spôsob, ako vo viacerých vrstvách realizovať spätnú väzbu a nastavovanie váhových stavov v procese učenia. Preto Minsky dokonca zavrhol celý koncept subsymbolickej AI. Až neskorší vývoj všeobecného algoritmu učenia sa, tzv. **back-propagation** algorithm znamenal skutočný prelom vo vývoji neurónových sietí, pričom v spojení s hlbokým učením sa tieto stali základom pre aktuálny veľký pokrok v oblasti umelej inteligencie.
NIELSEN, M. *Neural Networks and Deep Learning*. [on-line]. [cit. 19. januára 2022]. Dostupné na internete: <http://neuralnetworksanddeeplearning.com/>

114 Komplexnej analýze limitov a rizík súčasných systémov umelej inteligencie sa venujeme v 2. kapitole.

115 Nie je to problém pri systéme na rozpoznávanie číslic, ale môže to byť skutočný problém pri automatických bojových systémoch, autopilotov vozidiel, prostriedkov na podporu života, systémov na detekciu teroristických hrozieb, podporných systémov pre súdnictvo, atď.

Obhajcovia subsymbolických systémov umelej inteligencie – argumentujúc „architektúrou“ mozgu na úrovni neurónov – sú presvedčení, že prostriedkami symbolickej AI (symboly a logické vzťahy medzi nimi) nie je možné dosiahnuť umelú inteligenciu, keďže symboly v mozgu vznikajú na základe neurálnych procesov (na úrovni neurónov).[116]

I napriek všetkým úspechom, ktoré sme v súčasnosti vďaka subsymbolickým systémom, medzi ktoré patria neurónové siete, systémy hlbokého učenia atď., dosiahli, **dovolíme si s týmto názorom nesúhlasiť**, keďže stále nie sme schopní prekročiť hranice ANI, t.j. hranicu medzi úzko špecializovanými systémami (slabej) umelej inteligencie a všeobecnými systémami AGI.[117] Radi by sme uviedli aspoň dve výhrady...

Domnievame sa, že **ak budujeme systémy umelej inteligencie inšpirujúc sa ľudským mozgom, resp. sa chceme priblížiť analógii s mozgom, je treba vziať do úvahy metódy subsymbolickej i symbolickej AI**. Na jednej strane by tak išlo o postupy, ktoré sú jednou z možností ďalšieho rozvoja umelej inteligencie,[118] na druhej strane – čo je pre nás podstatné – by išlo o systémy, v ktorých na úrovni symbolickej AI by sme azda boli schopní ľahšie riešiť etické výzvy a implementovať pravidlá aj v prípade komplexných a pokročilých systémov AI.[119]

Aj keď – ako sme uviedli v kapitole 1.4. – logika prvého rádu sa nedokázala vysporiadať s neistotou a nedeterministickým vnímaním sveta, nesmieme ju vylúčiť z dizajnu systémov AI. V súčasnosti sa totiž robí opačný extrém – na základe preferovania subsymbolických systémov sa rezignovalo na logiku až do tej miery, že moderní vývojári umelej inteligencie o nej mnohokrát ani nechyrujú.[120] **Keďže sa však reálny svet skladá z objektov a vzťahov medzi nimi, logika prvého rádu patrí k fundamentom pri popise reálneho**

---

116 Por. MITCHELL, *Artificial Intelligence*, s. 31.

117 Informácie o ANI a AGI sú uvedené v kapitole 1.3.

118 V súčasnosti sa pokroky v moderných systémoch AI dosahujú (niekedy až radikálnym) vylepšovaním existujúcich algoritmov a súčasne kombináciou viacerých systémov, metód a postupov na dosiahnutie cieľa (viď poznámky č. 84 a 85).

119 Treba však uviesť, že doterajšie pokusy vytvoriť hybridný systém integrujúci subsymbolické i symbolické metódy AI nemali nejaký väčší úspech.
MITCHELL, *Artificial Intelligence*, s. 41.

120 Pre symbolické systémy postavené na logike sa dokonca zaužíval pejoratívny názov GOFAI – Good Old-Fashioned AI.
HAUGELAND J. *Artificial Intelligence: The Very Idea*. MIT Press, 1985.

sveta a ako taká by mala byť súčasťou návrhu sofistikovaných systémov AI.[121] [122]

Na základe doterajšieho výskumu v oblasti AI môžeme povedať, že systémy schopné zmysluplne poznávať veci musia mať kapacitu reprezentovať ich a uvažovať nad nimi porovnateľnú minimálne s logikou prvého rádu, pričom zatiaľ nevieme, akú presnú podobu budú tieto systémy mať. Ich logika totiž môže byť začlenená do pravdepodobnostných systémov uvažovania, do systémov hlbokého učenia alebo do nejakého hybridného dizajnu, ktorý ešte len bude vynájdený.[123]

Druhou výhradou je fakt, že i napriek veľkej snahe dnešné hlboké neurónové siete a algoritmy hlbokého učenia (teda dnešné subsymbolické systémy) len čiastočne simulujú mozgovú činnosť a prácu neurónov. **Biologické neuróny sa správajú inak ako uzly neurónových sietí a mozgové procesy sú v mnohom komplexnejšie**.

Uzol napodobňujúc činnosť neurónu vysiela impulz, resp. vo všeobecnosti číselné stavy. Výstupom biologického neurónu je však elektrický prúd spôsobený pohybom iónov v bunke. Jednotlivé výstupné prúdy sa prostredníctvom synáps posúvajú do ďalších buniek, kde zvyšujú ich membránový potenciál, ktorý je výsledkom nerovnovážnej koncentrácie iónov. Keď membránový potenciál neurónu prekročí určitú prahovú hodnotu, bunka vyšle impulz (používa sa termín spike), t. j. prúd, ktorý sa má odovzdať do ďalších buniek.[124] V klasických neurónových sieťach je tento proces simulovaný výstupným binárnym impulzom a váhovým stavom „synapsy“ spájajúcej dva uzly.

Biologické neuróny majú naviac vnútornú dynamiku, ktorá spôsobuje ich zmeny v čase.

121 Por. RUSSELL, *Human Compatible*, s. 271.

122 Rovnaký pohľad zdieľa i Demis Hassabis, zakladateľ a CEO spoločnosti DeepMind venujúcej sa najpokročilejším systémom AI: „Na dosiahnutie skutočnej inteligencie nestačia súčasné systémy hlbokého učenia, ktoré sú ekvivalentami zmyslových centier v mozgu, napr. vizuálneho alebo sluchového. Skutočná inteligencia je však oveľa viac než to, keďže je treba skombinovať zmyslové vstupy do myslenia vyššej úrovne a symbolického uvažovania. Musíme preto tieto systémy budovať až do symbolickej úrovne uvažovania a myslenia - t. j. do úrovne matematiky, jazyka a logiky.“
HEATH N. *Google DeepMind founder Demis Hassabis: Three truths about AI* [on-line]. TechRepublic, September 24, 2018. [cit. 14. augusta 2022].
Dostupné na internete: <https://www.techrepublic.com/article/google-deepmind-founder-demis-hassabis-three-truths-about-ai/>

123 RUSSELL, *Human Compatible*, s. 272.

124 *Action potential* [on-line]. [cit. 27. februára 2022].
Dostupné na internete: <https://en.wikipedia.org/wiki/Action_potential>

S plynutím času majú tendenciu vybíjať sa a znižovať svoj membránový potenciál, čoho dôsledkom je napríklad skutočnosť, že riedke impulzy na vstupe neurónu nespôsobia výstupný impulz![125]

Dôležitým rozdielom je i spôsob komunikácie medzi biologickými neurónmi a medzi uzlami. Biologické neuróny komunikujú a spracúvajú informácie asynchrónne, kým typické uzly neurónových sietí synchrónne, t.j. v jednom kroku všetky uzly jednej vrstvy neurónovej siete prečítajú vstup, vyrátajú výstup a posúvajú ho ďalej. U biologických neurónov je to iné – v každom okamihu môžu prijať vstupný signál a vytvoriť výstup bez ohľadu na správanie sa ostatných neurónov.[126]

Tiež treba poznamenať, že biologické systémy majú oveľa prepracovanejší systém spätnej väzby na úrovni neurónov (čo je súčasťou procesov učenia sa a adaptívnej činnosti), než sú dnešné metódy používané v neurónových sieťach (napr. back-propagation algoritmus spomínaný na začiatku tejto kapitoly).

Na jednej strane tak máme súčasné sofistikované neurónové siete, ktoré majú rastúce, ba až neúnosné výpočtové náklady (a tým aj vysokú spotrebu energie, časovú náročnosť, a pod.) potrebné na ich kvalitné vytrénovanie a rozvoj.[127] Tieto siete vedia byť veľmi efektívne v konkrétnych scenároch použitia, no ak by mali vedieť riešiť pre človeka jednoduché základné úlohy[128], resp. doterajšie zadania zovšeobecňovať a nebodaj aj chápať, sú v koncoch.

Na strane druhej máme prirodzenú inteligenciu s nepatrnou spotrebou energie, schopnú kreativity, riešenia problémov a multitaskingu. Ide o biologické systémy, ktoré si

125 KORAKOVOUNIS, D. *Spiking Neural Networks: where neuroscience meets artificial intelligence*. [on-line]. [cit. 27. februára 2022].
Dostupné na internete: <https://theaisummer.com/spiking-neural-networks/>

126 KORAKOVOUNIS, *Spiking Neural Networks: where neuroscience meets artificial intelligence*. [on-line]. [cit. 27. februára 2022].
Dostupné na internete: <https://theaisummer.com/spiking-neural-networks/>

127 „The cost of improvement is becoming unsustainable.“
THOMPSON, N. C., GREENEWALD, K., LEE, K., MANSO, G. F. *Deep learning computational cost*. [on-line]. [cit. 27. februára 2022].
Dostupné na internete: <https://spectrum.ieee.org/deep-learning-computational-cost>

128 „The easy things are hard.“ – táto problematika je rozoberaná v kapitole 1.9.

prirodzenou evolúciou osvojili spracovanie informácií a reagovanie na ne.[129] [130]

Ovocie pokračujúceho skúmania mozgových procesov a neurónov, snahy pochopiť, čo ich robí tak efektívnymi a odvaha aplikovať tieto zistenia do oblasti umelej inteligencie viedli k vzniku tretej generácie neurónových sietí, tzv. **spiking neural networks** (SNNs) a ich implementácie v rámci neuromorfnej hardvérovej architektúry.[131] [132]

Vrátiac sa k začiatku tejto diskusie, t.j. rozporovaniu tvrdenia, že súčasné algoritmy strojového učenia sú schopné dosiahnuť skutočnú umelú inteligenciu, vidíme, že je pred nami ešte dlhá cesta a vývoj, ktorý v kontexte doterajších subsymbolických i symbolických prístupov môže priniesť veľa zaujímavých poznatkov a progresu na ceste (nielen) k uvedomelej umelej inteligencii.

1.8. Striedanie ročných období a najbližšia predpoveď počasia

Skúsme spraviť malú retrospektívu vývoja umelej inteligencie. Roky po Dartmouthskom seminári ubiehali a veľmi optimistické predpovede o pokroku vo vývoji systémov umelej inteligencie sa nenapĺňali.[133] S vyčerpaním potenciálu konkrétnych dobových poznatkov v oblasti AI a schopností dobovej techniky tak veľmi rýchlo prichádzalo k strate nadšenia i utlmeniu investícií do tejto oblasti.

Vývoj bol udržiavaný v zásade len v rámci základného výskumu viacerých univerzitných

129 KORAKOVOUNIS, *Spiking Neural Networks: where neuroscience meets artificial intelligence*. [on-line]. [cit. 27. februára 2022].
Dostupné na internete: <https://theaisummer.com/spiking-neural-networks/>

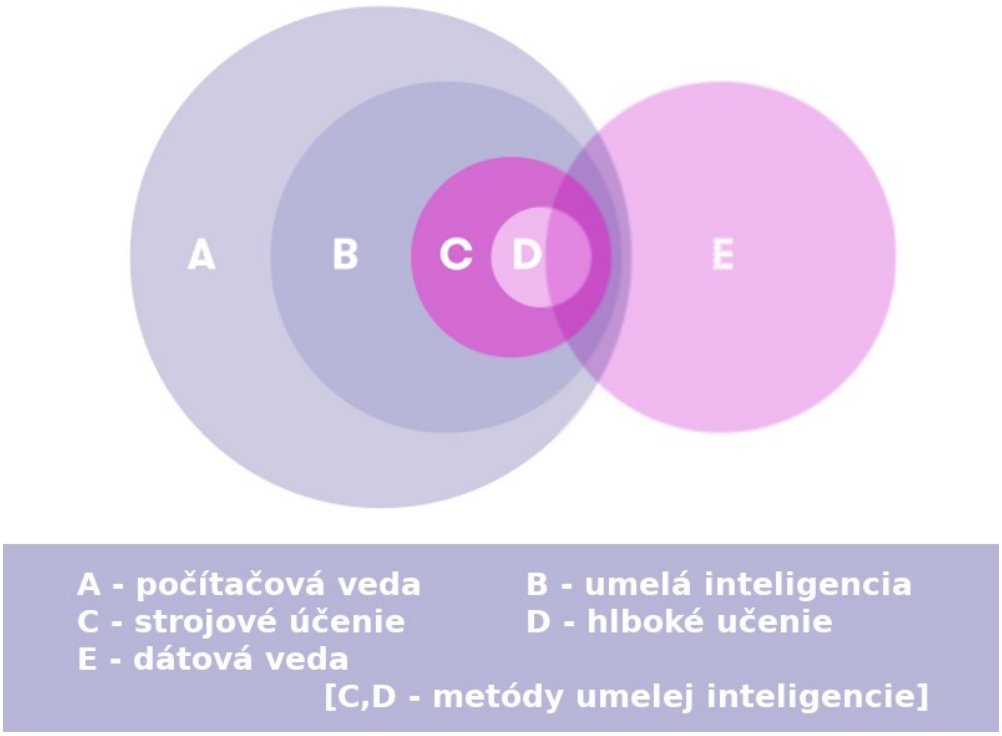
130 V našom kontexte sú tieto systémy i súčasťou etického rámca, ktorý biológiu presahuje.

131 PFEIFFER, M., PFEIL, T. *Deep Learning With Spiking Neurons: Opportunities and Challenges*. [on-line]. [cit. 27. februára 2022].
Dostupné na internete: <https://www.frontiersin.org/articles/10.3389/fnins.2018.00774/full>

132 Predchádzajúce generácie neurónových sietí sú založené na McCullochových Pittsových neurónoch (t.j. prahových hradlách), resp. sigmoidálnych hradlách, pričom sa javí, že jeden spike neurón dokáže riešiť takú konkrétnu biologickú funkcionalitu, ktorá by si vyžadovala stovky skrytých jednotiek sigmoidálnej neurónovej siete.
MAASS, W. *Networks of spiking neurons: The third generation of neural network models*. [on-line]. [cit. 28. februára 2022].
Dostupné na internete: <https://www.sciencedirect.com/science/article/abs/pii/S0893608097000117>

133 Viac o optimistických predpovediach a predikciách hlavných protagonistov je uvedené v poznámke č. 68.

centier a pokroku v príbuzných oblastiach (napr. robotika a kybernetika, výpočtová technika, dátová a počítačová veda, v rámci priemyslu a vojenského vývoja), z čoho ťažila aj veľmi vágne ohraničená oblasť umelej inteligencie, ako sme uviedli na konci kapitoly 1.2. Vzťah medzi dátovou vedou, počítačovou vedou a umelou inteligenciou ako vedným odborom vyjadruje schematický obrázok č. 5.



Obr. 5: Vzťah medzi dátovou vedou, počítačovou vedou a umelou inteligenciou.

I napriek útlmu ovocím základného výskumu a pokroku v príbuzných oblastiach bol posun v hľadaní nových metód a prístupov k riešeniu problémov umelej inteligencie. Každý väčší posun sa následne stával štartérom nového entuziazmu a záujmu o umelú inteligenciu, prichádzali nové pozitívne predikcie, ktorých ovocím bol i ďalší prísun investícií a podpory pre túto oblasť. A po vyčerpaní potenciálu daného posunu znovu nastával útlm...

Toto – až do poslednej dekády viac menej pravidelné – **striedanie aktívneho rozvoja a útlmu v oblasti umelej inteligencie sa nazýva aj striedanie ročných období**:

- jar umelej inteligencie (AI spring) – obdobie prekotného rozvoja, ktoré štartuje novými ideami a pokrokom vo výskume, čoho ovocím sú predikcie prevratného pokroku vo vývoji systémov AI častokrát sprevádzané mediálnym boomom a následným prílevom štátnych dotácií i rizikového kapitálu do akademického bádania i komerčných startupov.

- zima umelej inteligencie (AI winter) – prevratný pokrok neprichádza, výsledky síce sú, no vôbec nespĺňajú extrémne očakávania. Prílev financií a kapitálu ustáva, startupy krachujú a akademický výskum sa spomaľuje a obmedzuje.

Uvedený vzor sa vo vývoji umelej inteligencie v rozličných podobách opakoval v päť až desaťročných cykloch.[134]

S veľkým pokrokom vo vývoji algoritmov vychádzajúcich zo štatistických a pravdepodobnostných teórií, ktoré umožnili rozvoj strojového učenia, v súčasnosti postupne prišlo – až na niekoľko špecifických oblastí – k odmietnutiu symbolických systémov AI[135] a veľkému nástupu neurónových sietí a strojového učenia. V posledných dvoch desaťročiach tak striedanie ročných období reálne prebieha už len v rámci vlastných cyklov vývoja subsymbolických systémov AI, keďže viackrát sa ukázalo, že na riešenie skutočných problémov reálneho sveta sú aj moderné systémy strojového učenia prikrátke...

V poslednej dekáde možno diskutovať o ďalšom vývoji umelej inteligencie i z pohľadu týchto vývojových cyklov.

Umelá inteligencia v súčasnosti zažíva nebývalý rozmach, dynamický vývoj, extrémny rozsah praktického aplikovania jej systémov, špičkové zabezpečenie výskumu a vývoja, spoločenskú akceptáciu i prehnané mediálne pokrytie a očakávania. Vyzerá to skoro ako definícia jarnej fázy AI.

Na druhej strane však treba povedať, že aktuálne veľké pokroky nie sú dôsledkom revolučných zmien v AI, ale len evolučného vývoja existujúcich metód a ich sofistikovaných kombinácií, technologickej podpory z iných oblastí (hlavne rozvoj informačných a komunikačných technológií), dostupnosť extrémne veľkých datasetov, atď. Prevratný pokrok však v podstatných aspektoch neprichádza (vytvorenie AGI). To niekomu skôr pripomenie prvé zimné mrazy...

Keďže miera akceptácie a adaptácie systémov AI v modernej spoločnosti presiahla

134 Por. MITCHELL, *Artificial Intelligence*, s. 34.

Autorka dokonca spomína, že v čase ukončenia jej vysokoškolského štúdia (1990) sa vývoj AI nachádzal tak hlboko „v zimnom období“, až je radili, aby do profesného životopisu v rámci žiadostí o zamestnanie umelú inteligenciu vôbec neuvádzala.

135 Odmietnutie bolo často sprevádzané dešpektom zhmotneným do označenia symbolickej AI ako GOFAI (podľa Douglasa Hofstadtera ide o skratku pre good old old-fashioned AI).

hranicu, za ktorou je ťažké si predstaviť návrat späť (resp. zmenu priorít rozvoja vedomostnej spoločnosti a Industry 4.0) a keďže trochu plochá krivka základného výskumu a vývoja systémov AI je súčasťou permanentného technologického pokroku, ktorý celkovo v oblasti rozvoja vedy, techniky, objemu dát a znalostí dosahuje takmer exponenciálny rast,[136] dovolíme si predpokladať, že na aktuálny a budúci vývoj v oblasti umelej inteligencie už nebude možné v plnej miere aplikovať striedanie jarných a zimných období. Výskum a vývoj systémov AI bude neustále pod tlakom prakticky všetkých spoločenských oblastí, v ktorých sa AI v súčasnosti využíva.

Pozorný čitateľ určite vníma, že spochybňujúc pokračovanie vývojových cyklov sme skĺzli do oblasti ANI, t.j. k úzko špecializovaným systémom (slabej) umelej inteligencie, ktorým sa v súčasnosti neskutočne darí.

Inak to však môže byť s AGI, t.j. skutočnou umelou inteligenciou, ktorá sa stáva definitívnou métou súčasného základného výskumu v oblasti AI.[137] Tu si vývoj nedovolíme predpovedať, keďže na jednej strane principiálny prelom vo vývoji metód umelej inteligencie nenastáva, na strane druhej však vylepšovanie a kombinácia súčasných techník, extrémne veľké datasety, výpočtové systémy s výkonom dizajnovaným priamo pre aplikácie umelej inteligencie i všeobecný vedomostný a výskumný potenciál môžu veľmi ľahko viesť aj k prípadnej skokovej zmene a vývojovému prelomu.

V rámci viac než šesťdesiatich rokov bádania a snáh o vytvorenie všeobecnej umelej inteligencie mnohí vedci a výskumníci viackrát takmer opustili ideál AGI. V súčasnosti – keď sa debata o umelej inteligencii stala súčasťou hlavného spoločenského prúdu – sa mnohým tento ideál javí ako samozrejmý. Predpokladáme, že len niekoľko rokov bude stačiť, aby sme videli, ktoré tendencie prevažujú: či optimizmus pretrváva, alebo fenomén AGI sa v rámci pokračujúceho výskumu bude javiť ako oveľa sofistikovanejší,

136 ŠANTAVÝ, *Analýza virtuálneho sveta v kontexte náboženskej formácie, pôsobenia a života Cirkvi,* [online], s. 20, 25, 31, 34. [cit. 7. januára 2022]. Dostupné na internete: <https://peter.santavy.cloud/data/uploads/docs/analyza_virtualneho_sveta_v_kontexte_nabozenskej_formacie_posobenia_a_zivota_cirkvi.pdf>

137 Filozofi umelej inteligencie, Vincent Müller a Nick Bostrom zverejnili v roku 2013 prieskum vykonaný medzi výskumníkmi AI, v ktorom mnohí vyjadrili 50% šancu na dosiahnutie umelej inteligencie na úrovni človeka do roku 2040.
MÜLLER, V. C., BOSTROM, N. *Future Progress in Artificial Intelligence: A Survey of Expert Opinion*. In: *Fundamental Issues of Artificial Intelligence. Cham.* Switzerland: Springer International, 2016, 555-72.

komplexnejší a náročnejší nielen na realizáciu, ale hlavne na uchopenie a pochopenie.

1.9. Jednoduché veci sú ťažké

Nadväzujúc na predchádzajúcu kapitolu treba povedať, že v určitej miere nepôjde o nič nové, keďže súčasťou každej zimy umelej inteligencie bolo zistenie, akým problémom pre systémy AI je schopnosť realizácie niektorých vecí. Päťdesiat rokov od konferencie v Dartmouthe to John McCarthy vyjadril slovami: „AI bola ťažšia, ako sme si mysleli“.[138] A Marvin Minsky už oveľa skôr poznamenáva, že výskum v oblasti umelej inteligencie odkryl **dôležitý paradox: „Jednoduché veci sú ťažké“[139], keďže riešenie úloh, ktoré sa človeku zdajú byť jednoduché, je pre umelú inteligenciu veľmi náročné. A naopak to, čo je pre človeka extrémne náročné, dokážu systémy AI zvládať veľmi dobre.[140]**

Pôvodné ciele vývoja umelej inteligencie – počítače, ktoré dokážu s nami komunikovať ľudskou rečou, vedia popísať, čo „vidia“ kamerou, „chápu“ koncepty na základe len niekoľkých príkladov – to všetko sú veci, ktoré malé dieťa poľahky zvláda, no pre AI sú oveľa ťažšie, ako napr. komplexná diagnostika choroby, víťazstvo v šachu alebo v Go nad najlepšími ľudskými šampiónmi alebo riešenie sofistikovaného algebraického problému.[141] [142]

Minský to vyjadril slovami: „Vo všeobecnosti si najmenej uvedomujeme to, čo naša myseľ robí najlepšie.“[143] A Sparsh Chadha dodáva: „Stroje ... dokážu vykonávať len úlohy, na ktoré boli vytrénované. Vďaka našej schopnosti dynamického myslenia a zdravému

138 MOEWES, C., NÜRNBERGER, A. *Computational Intelligence in Intelligent Data Analysis*. New York: Springer, 2013, s. 135.

139 MINSKY, M. *Society of Mind*. New York: Simon & Schuster, 1986, s. 29.

140 Kognitívny vedec Steven Pinker to vyjadruje celé: **„The hard things are easy, but the easy things are hard.“**
TROTT, D. *The hard things are easy, but the easy things are hard*. [on-line]. [cit. 8. januára 2022]. Dostupné na internete: <https://www.campaignlive.com/article/hard-things-easy-easy-things-hard/1498154>

141 MITCHELL, *Artificial Intelligence*, s. 33-34.

142 Ide o tzv. Moravcov paradox, ktorý bol jedným z výsledkov výskumu Hansa Moravca, Rodney Brooksa a Marvina Minského v osemdesiatych rokoch.
MORAVEC, H. *Mind Children: The Future of Robot and Human Intelligence*. Cambridge, Mass.: Hardvard University Press, 1988.

143 MINSKY, *Society of Mind*, s. 29.

rozumu výrazne prevyšujeme stroje. Bez ohľadu na to, ako veľmi natrénujeme naše modely, na koľkých tréningových cykloch trénujeme naše stroje, stále existuje priestor neistoty, ktorý by sa dal človekom veľmi jednoducho vyriešiť, ale umelá inteligencia by zlyhala, pretože jej chýba schopnosť zdravého rozumu.“[144]

Ide o skúsenosť sprevádzajúcu celé desaťročia vývoja umelej inteligencie, ktorá nám pripomína, ako komplexná a subtílna je ľudská myseľ a čo všetko je vo vývoji AGI ešte pred nami.

Jednou z podmienok zvládnutia pre človeka jednoduchých, ale pre AI náročných vecí je prechod od systémov, ktoré sa javia ako inteligentné (napr. systém AI na jazykový preklad, systém na popis objektov zosnímanej scény a pod.) **k systémom, ktoré sú inteligentné** (AI systém, ktorý rozumie prekladanému textu, systém, ktorý chápe obsah zosnímanej scény a pod.).[145] Avšak rozumieť a chápať súvislosti - to sa znovu dostávame do oblasti všeobecnej a silnej umelej inteligencie (AGI).

Prakticky vo všetkých strategických oblastiach rozvoja umelej inteligencie je AGI túženým cieľom, keďže jej schopnosti by umožnili riešiť problémy na úplne inej úrovni. V kombinácii s technologickými možnosťami súčasnej informatizácie a automatizácie sa možnosti AGI javia ako úžasné.

Napr. analýza cestnej premávky na základe všetkých dostupných senzorov autonómneho vozidla – teda schopnosť ľudského komplexného vnímania, ktorá by bola spojená s vizuálnymi, zvukovými, radarovými, lidarovými, atď. senzormi a zároveň by umožňovala extrémne rýchle reakcie v priamej interakcii s elektronickým ovládaním vozidla.[146]

Tieto predpokladané úžasné možnosti však idú ruka v ruke s výzvami, ktoré sa doteraz v oblasti umelej inteligencie neriešili – etický rozmer činnosti strojov: ako ho definovať, ako ho realizovať, ako fixovať požadované vlastnosti a ako ich

144 CHADHA, S. *“Common Sense” is the Dark Matter of Artificial Intelligence*. [on-line]. [cit. 4. februára 2022].
Dostupné na internete: <https://hackernoon.com/the-dark-matter-of-ai-common-sense-is-not-so-common>

145 Pozri poznámku č. 56 a súvisiaci text v kapitole 1.1.

146 Súčasné autonómne systémy sú mnohokrát veľmi pomalé, ak sa majú rozhodovať napr. v situácii zložitej križovatky.

aplikovať do predmetu činnosti daného systému AI.[147]

1.10. Transhumanizmus a umelá inteligencia – tandem i súperi

Koketovanie s myšlienkou AGI a túžba po vytvorení uvedomelej umelej inteligencie je často konfrontovaná s porovnávaním s inteligenciou človeka. V optike problémov, ktoré vývoj AGI obnáša a v obavách z disproporcie medzi uvedomelou umelou a ľudskou inteligenciou, mnoho vedcov a futurológov sa zaoberá myšlienkami transhumanizmu, t.j. futurologickým a filozofickým konceptom, ktorý pojednáva o možnostiach transformácie človeka prostredníctvom využitia moderných technológií. Je to pohľad do budúcnosti, ktorý je založený na premise, že ľudia vo svojej súčasnej podobe nereprezentujú koniec vývoja, ale iba jeho ranú fázu.[148]

V zásade ide o dva základné prieniky medzi umelou inteligenciou a transhumanizmom:

- integrácia systémov AI a človeka
- existenciálne riziko vyplývajúce z pokročilej umelej inteligencie

Zavádzanie prvkov umelej inteligencie obnáša aj asistenčné technológie, ktoré zvyšujú kvalitu života, od tých jednoduchých, napr. implementácia prvkov rozšírenej reality (augmented reality) v rámci rôznych nositeľných zariadení (wearables), až po oveľa sofistikovanejšie, napr. pripojenie robotických končatín na nervovú sústavu

147 Etika systémov umelej inteligencie je v súčasnosti pomerne zavedená oblasť výskumu AI. U AGI však neide o etiku vývoja, nasadenia, používaných dát, atď., ale priamo o etický rozmer činnosti všeobecnej umelej inteligencie.

148 Transhumanizmus je futurologický a filozofický koncept, ktorý sa zaoberá možnosťami transformácie človeka prostredníctvom využitia moderných technológií. Ide o filozofické a intelektuálne hnutie, ktoré obhajuje zlepšenie ľudského stavu vývojom a sprístupnením sofistikovaných technológií, ktoré dokážu výrazne zvýšiť dĺžku a kvalitu života, kognitívne schopnosti. Transhumanistickí myslitelia skúmajú potenciálne výhody a nebezpečenstvá vznikajúcich technológií, ktoré by mohli prekonať základné ľudské obmedzenia, ako aj etiku používania takýchto technológií.
MERCER, C., TROTHEN, T. J. *Religion and Transhumanism: The Unknown Future of Human Enhancement*. Praeger, 2014.
Niektorí transhumanisti sa domnievajú, že ľudia sa nakoniec budú môcť premeniť na bytosti so schopnosťami tak výrazne rozšírenými oproti súčasnému stavu, že si zaslúžia označenie post-humánne bytosti (pojednáva o tom samostatný komplexný futuristický smer označovaný ako posthumanizmus).

hendikepovaného človeka[149]. V tomto kontexte sa pre časť výskumu javí atraktívnou myšlienka prepojiť vhodné systémy z umelej inteligencie s telom a nervovou sústavou človeka. Osobitne, ak na jedne strane dosahujeme pomerne veľké úspechy vo vývoji ANI, no súčasne sa len málo posúvame k dosiahnutiu méty AGI.

Keďže jednou z tém a oblastí transhumanistického výskumu je snaha ochrániť ľudstvo pred existenčnými rizikami,[150] ako je napríklad jadrová vojna alebo zrážka s asteroidom, smerovanie k všeobecnej a uvedomelej umelej inteligencii spojené s rizikami, ktoré jej potenciál môže obnášať, viedli transhumanistickú scénu aj k vnímaniu existenciálnych rizík vyplývajúcich z pokročilej umelej inteligencie.[151]

Ak by sme len v úzkom kontexte tejto publikácie chceli uchopiť etické výzvy spojené s transhumanizmom[152], išlo by buď o problémy súvisiace s vplyvom virtuálneho sveta na život človeka a spoločnosti, ktoré sme popísali v 2. kapitole našej licenciátskej práce Analýza virtuálneho sveta v kontexte náboženskej formácie, pôsobenia a života Cirkvi[153], teda problémy týkajúce sa slabej umelej inteligencie, ktoré sú analogické problémom

149 Elon Musk v rámci svojich vizionárskych aktivít (Neuraling Corporation) vytvára rozhranie medzi mozgom a počítačom pre rozšírenie možností človeka, ako napr. integrácia robotických končatín, alebo prepojenie ľudského mozgu a umelej inteligencie pre dosiahnutie symbiózy, ktorú by sme poľahky mohli zaradiť do oblasti transhumanizmu.
Neuralink [on-line]. [cit. 5. augusta 2020].
Dostupné na internete: <https://en.wikipedia.org/wiki/Neuralink>

150 „Elon Musk už dlho hovorí, že umelá inteligencia bude musieť ľudské schopnosti skôr rozširovať, než s nimi súťažiť, aby sa vyhla hrozivej budúcnosti.“
FERRIS, R. *Elon Musk thinks we will have to use AI this way to avoid a catastrophic future*. [on-line]. [cit. 24. januára 2022].
Dostupné na internete: <https://www.cnbc.com/2017/01/31/elon-musk-thinks-we-will-have-to-use-ai-this-way-to-avoid-a-catastrophic-future.html>

151 BOSTROM, N. *A history of transhumanist thought*. In: *Journal of Evolution and Technology.* [on-line]. 2005, zv. 14, vyd. 1. [cit. 8. januára 2022].
Dostupné na internete: <http://www.nickbostrom.com/papers/history.pdf>

152 Teda etické výzvy spojené s umelou inteligenciou, nie celý rozsah problémov týkajúcich sa transhumanizmu.

153 ŠANTAVÝ, *Analýza virtuálneho sveta v kontexte náboženskej formácie, pôsobenia a života Cirkvi,* [on-line], s. 25-47. [cit. 8. januára 2022]. Dostupné na internete: <https://peter.santavy.cloud/data/uploads/docs/analyza_virtualneho_sveta_v_kontexte_nabozenskej_formacie_posobenia_a_zivota_cirkvi.pdf>

s ostatnými aspektami informačných technológií, kybernetického priestoru a virtuálneho sveta. Alebo by išlo – lepšie povedané pôjde – o problémy oscilujúce medzi výzvami spojenými s etikou uvedomelej umelej inteligencie a psychológiou, morálnymi aspektami i vierou postmoderného človeka.

Pri masívnom prieniku systémov umelej inteligencie do reálneho života treba teda rátať s vplyvom transhumanistickej filozofie, ako napr. s mravným redukcionizmom a relativizmom, falošnou premisou, že technické „vylepšovanie“ človeka v kontexte umelej inteligencie vedie k jeho kultúrnemu i morálnemu zlepšovaniu[154], materialistickému a technokratickému chápaniu šťastia a naplnenia[155], falošným rovnostárstvom s odovzdaním vlády superpočítačom (~umelej inteligencii)[156] a pokriveným vnímaním eschatologickej roviny viery v kontexte transfigurizmu, t.j. náboženského transhumanizmu.[157]

1.11. Súčasnosť – umelá inteligencia „na koni“

V súčasnosti sme svedkami enormného rozmachu vo vývoji, tvorbe, realizácii, nasadení, poskytovaní i využívaní technológií umelej inteligencie. Tento rozmach stojí na viacerých pilieroch, ktoré sú ovocím súčasného rozvoja informačnej spoločnosti:

- pokrok vo vývoji metód a systémov umelej inteligencie, osobitne v oblasti neurónových sietí;
- existencia, jednoduché získavanie a bezprecedentná dostupnosť[158] extrémneho

154 STRAHOVNIK, V. *Virtues and transhumanist human enhancement*. In: PETROUŠEK R., ŽALEC B., eds. *Transhumanism as a Chalange for Ethics and Religion*. 2021, s. 37 – 44.

155 *Transhumanist Declaration*. [on-line]. [cit. 26. januára 2022]. Dostupné na internete: <https://hpluspedia.org/wiki/Transhumanist_Declaration>

156 LEE, I. *Equalism: Paradise Regained*. In: LEE N. (ed.). *The Transhumanist Handbook*. Springer, 2019, s. 849 – 863.

157 Solídny náhľad do tejto problematiky v optike kresťanského svetonázoru a Katolíckej cirkvi uvádza Dr. Vivoda v svojom článku Transhumanizmus a Katolícka Cirkev.
VIVODA, M. *Transhumanizmus a Katolícka Cirkev*. In: *Nové Horizonty*. 2021, roč. 15, č. 3, s. 121-128. ISSN: 1337-6535, EAN 977133765400605

158 Už v roku 2010 Erich Smid, v tom čase CEO Google, na konferencii Techonomy v kalifornskom Lake Tahoe vyhlásil: „V súčasnosti každé dva dni vyprodukujeme toľko informácií, koľko sme vytvorili od úsvitu civilizácie až do roku 2003.“
SIEGLER, M. G. *Eric Schmidt: Every 2 Days We Create As Much Information As We Did Up To 2003*

množstva štruktúrovaných i neštruktúrovaných dát;[159]

- moderné prístupy, ktoré vyžadujú adaptívne procesy schopné spracovávať také kvantum dát;
- potreba automatizácie – rozvoj automatizácie smerujúci k potrebe autonómnych systémov schopných samostatnej adaptívnej činnosti pri spracúvaní veľkého množstva neštruktúrovaných dát v konkrétnej oblasti.

Tieto piliere sú súčasťou tvoriacej sa a (v niektorých častiach sveta) rozvíjajúcej sa informačnej a znalostnej spoločnosti, ktorá je sprevádzaná paradigmatickou zmenou: informačné technológie a dáta sa z roviny prostriedku dostávajú do roviny kontextu spoločnosti až do tej miery, že na základe zmien v technológiách a vzhľadom na prevratný vedecký i technologický pokrok sa mení medziľudská komunikácia vo svojej podstate, menia sa vzťahy, mení sa spoločnosť a princípy jej fungovania.[160]

Nástup a rozšírenie systémov umelej inteligencie patrí k tejto zmene paradigmy, pričom si dovolíme tvrdiť, že smerovanie k AGI a dosiahnutie systémov uvedomelej umelej inteligencie by pre mnohých bolo jej naplnením.

[on-line]. [cit. 12. decembra 2021].
Dostupné na internete: <http://techcrunch.com/2010/08/04/schmidt-data/>

159 Senzorické systémy, počítačové siete, digitalizácia, informatizácia vedy i technologických procesov a predovšetkým on-line priestor komunikácie a sociálnych sietí sú zdrojom veľkého množstva dát. „Big data“, ako súčasť technologickej a informačnej revolúcie, umožnili rozvoj viacerých v súčasnosti dominantných metód umelej inteligencie, keďže vďaka veľkému množstvu dát je možné dostatočne a relevantne trénovať systémy AI, aby boli použiteľné v reálnom nasadení.
Por. MITCHELL, *Artificial Intelligence*, s. 80.

160 ŠANTAVÝ, *Analýza virtuálneho sveta v kontexte náboženskej formácie, pôsobenia a života Cirkvi,* [on-line], s. 19-20. [cit. 7. decembra 2021].
Dostupné na internete:
<https://peter.santavy.cloud/data/uploads/docs/analyza_virtualneho_sveta_v_kontexte_nabozenskej_formacie_posobenia_a_zivota_cirkvi.pdf>

2. Limity a riziká súčasných systémov umelej inteligencie

Každá dostatočne pokročilá technológia je na nerozoznanie od mágie.[161]

V súčasnosti sme svedkami rozsiahleho nasadzovania systémov umelej inteligencie do praxe, pričom nejde len o širokospektrálne využitie v oblasti vedy a výskumu, resp. v niektorých strategických oblastiach, ale aj o bežné používanie v spotrebnej elektronike a v systémoch, ktoré sú súčasťou nášho každodenného života. Využitie systémov AI sa naviac stáva tak samozrejmým, že ich bežné používanie si ani neuvedomujeme, avšak ich akceptujeme a pomaly – zvykajúc si na benefity ich nasadenia – ich až vyžadujeme, keďže v mnohom nám prinášajú pridanú hodnotu, ktorej sa nechcem vzdať.[162]

Ak použijeme analógiu z oblasti kybernetickej bezpečnosti, zameranie nastupujúcich generácií na komfort a benefity informačných technológií je mnohokrát silnejšie, než opatrnosť a bezpečné správanie sa v rámci ochrany pred kybernetickými hrozbami, únikom a zneužitím osobných údajov a pod. Zámerne spomínam nastupujúcu generáciu, ktorá sa hrdí familiárnosťou vo využívaní informačných technológií a médií, v rámci vzdelávania dostáva aspoň základy gramotnosti v oblasti informačnej bezpečnosti, ale návykovosť a komfort mnohých technológií v reálnom použití víťazí nad zdravou mierou opatrnosti a aplikovania bezpečnostných zásad.

Je takmer isté, že podobná ľahkovážnosť bude sprevádzať spoločnosť aj pri využívaní systémov umelej inteligencie. Súčasná miera akceptácie jednoduchých systémov AI a spôsob ich využívania dáva tušiť, že používanie sofistikovaných systémov AI môže obnášať riziká, ktoré si ani nevieme predstaviť.

V tejto kapitole sa však venujme problémom, o ktorých vieme a treba s nimi rátať pri návrhu, tvorbe a využívaní systémov umelej inteligencie. Stále sa pritom pohybujeme v oblasti ANI, teda úzko špecializovaných systémov umelej inteligencie

161 Arthur C. Clarke.

162 Napr. pre zákazníkov to môže byť využitie jednoduchých systémov AI vo fotoaparátoch chytrých telefónov – modely, ktoré tieto funkcie nemajú, sú v danej kategórii prakticky nepredateľné.
Pre majiteľov elektronických obchodov sa stali nepostrádateľnými systémy, ktoré dokážu vytvárať profily prichádzajúcich návštevníkov e-shopov a zvýšiť šancu predaja rôznych produktov.
Viaceré bankové domy si v súčasnosti nevedia predstaviť niektoré svoje on-line služby bez dohľadu natrénovanej umelej inteligencie, ktorá dokáže odhaľovať podvody a falošné transakcie...

(narrow AI), ktoré sú optimalizované na zvládnutie konkrétnej úlohy, resp. množiny úloh. Ide súčasne o systémy slabej umelej inteligencie (weak AI), ktoré vykazujú inteligentné správanie na základe modelov a aplikovaných metód i tréningových dát. Hovoríme teda o systémoch, ktoré sú zamerané na riešenie konkrétnych úloh a sú závislé na ľudskom vstupe a konfigurácii.[163]

Keďže v súčasnosti neexistujú systémy AGI, t.j. silnej a všeobecnej umelej inteligencie, zameranie na ANI je samozrejmé. Naviac – vzhľadom na aktuálnu absolútnu preferenciu neurónových sietí a strojového učenia – sa sústredíme na ich súčasnú prezentáciu hlbokými neurónovými sieťami (deep neural networks) a hlbokým učením (deep learning). Pokúsime sa poukázať na limity a riziká týchto systémov AI, popísať technologické výzvy v oblasti bezpečnosti procesov umelej inteligencie a s tým súvisiacej kybernetickej bezpečnosti, uvedomiť si extrémnu komplexnosť týchto systémov a akcentovať viaceré dôsledky, ktorých zneužitie môže znamenať vážne ohrozenie pre jednotlivcov i spoločnosť.

2.1. Vybrané limity a rizikové faktory systémov umelej inteligencie

Pri vývoji, realizácii a nasadzovaní prvkov AI je treba rátať s viacerými obmedzeniami a rizikami súčasných systémov umelej inteligencie.

Ponajprv malé prirovnanie – ak vnímame neurónové siete a subsymbolickú AI ako systémy hlboko inšpirované činnosťou mozgu na úrovni neurónov[164], pre zdôraznenie **dôsledkov zlyhania prvkov neurónových sietí pre ich celkovú funkčnosť** môžeme použiť veľmi hrubú analógiu s mozgovými poruchami spôsobenými na bunkovej úrovni (neurón ~ uzol neurónovej siete), ktoré môžu viesť k veľmi vážnym psychickým poruchám.

Napríklad serotonín (5-hydroxytryptamin, skratka 5-HT) je biologicky aktívna molekula, ktorá pôsobí ako dôležitý neurotransmiter prenášajúci vzruchy medzi neurónmi. Aby po uvoľnení z jednej bunky mohol prenášať vzruchy na inú, viaže sa na príslušné povrchové receptory, medzi ktoré patrí aj serotonínový receptor 2A (5-HT_{2A})[165]. Nesprávna hladina a funkcia receptorov 5-HT_{2A} súvisí s vážnymi psychickými poruchami, vrátane

163 Viac o úzko špecializovaných a slabých systémoch AI pojednáva kapitola 1.3.

164 O systémoch AI inšpirovaných činnosťou mozgu na úrovni neurónov sme pojednávali v kapitole 1.5.

165 *5-HT_{2A} receptor* [on-line]. [cit. 20. januára 2022].
Dostupné na internete: <https://en.wikipedia.org/wiki/5-HT2A_receptor>

bipolárnej poruchy, ťažkých depresií a schizofrénie. Správnu (zdravú) rovnováhu serotonínových receptorov v mozgu môže poškodiť veľa faktorov, napr. drogy, niektoré choroby, neurologické lieky, ale aj nedostatok spánku.[166]

Podobne aj u neurónových sietí je veľa faktorov súvisiacich nielen s ich dizajnom a technologickým riešením, ale aj s konkrétnymi zvolenými parametrami, trénovacím datasetom, vyhodnocovaním a pod., ktoré môžu ovplyvniť ich správnu a bezpečnú funkčnosť.

Túto analógiu akcentuje aj skutočnosť, že súčasné systémy hlbokého učenia sú priamo modelované na základe objavov a poznatkov z neurovedy – inak povedané – moderné hlboké neurónové siete napodobňujú zodpovedajúce štruktúry mozgu. Napríklad v súčasnosti najúspešnejšie systémy AI pre spracovanie obrazu sú hlboké siete, ktorých štruktúra napodobňuje časti vizuálneho systému mozgu.[167] [168]

2.1.1. Neurónová sieť ako „black box“

Jedno zo základných rizík vyplýva zo skutočnosti, že **prakticky nevieme, na základe čoho robia hlboké neurónové siete svoje rozhodnutia**.[169] Vieme, ako nadizajnovať neurónovú sieť pre konkrétnu oblasť použitia. Vieme, ako ju natrénovať a v rámci možností aj otestovať. Keďže však neurónová sieť neobsahuje súbor presných softvérových postupov na úrovni logického myslenia, ale je tvorená len stohom rovníc, len húštinou ťažko interpretovateľných operácií s číslami, ktoré fungujú na základe správneho nastavenia váh, konštánt a prahových hodnôt, **v zásade nevieme, čo presne sa neurónová sieť naučila a ako spoľahlivo to dokáže aplikovať** nielen v bežnej prevádzke, ale osobitne v hraničných situáciách za extrémnych podmienok na vstupe, či

166 GREGOROVÁ, D. *Nedostatek spánku působí na mozek* [on-line]. [cit. 20. januára 2022]. Dostupné na internete: <https://www.osel.cz/12127-nedostatek-spanku-pusobi-na-mozek.html>

167 MITCHELL, *Artificial Intelligence*, s. 71.

168 Ide o konvolučné neurónové siete (convolutional neural networks, skratky ConvNets alebo CNN), ktoré sú dizajnované podľa vizuálneho cortexu v mozgu. Pomenovanie „konvolučné“ sa odvíja od názvu matematickej operácie vykonávanej nad vstupnými hodnotami recepčných polí v rámci jednotlivých aktivačných máp skrytých vrstiev neurónovej siete.
Por. MITCHELL, *Artificial Intelligence*, s. 72-80.
KOUSHIK, J. *Understanding Convolutional Neural Networks*. [on-line]. [cit. 31. januára 2022]. Dostupné na internete: <https://arxiv.org/abs/1605.09081>

169 Por. MITCHELL, *Artificial Intelligence*, s. 39.

pri činnosti systému. Táto miera nevedomosti rastie s mierou komplexnosti neurónovej siete, t.j. počtom skrytých vrstiev a uzlov, použitými algoritmami a prítomnosťou špecifických úprav.[170]

Už od školských lavíc poznáme otázku – ukáž, ako si to spravil – na ktorú odpovedáme ozrejmením postupu, akým sme prišli k výsledku. Vyjadrenie postupu tak umožňuje nielen jednu z možností verifikácie správnosti výsledku, ale i overenie postupu, ktorým k výsledku prichádzame. Ide tak o jeden zo základných stavebných kameňov dôvery a istôt, na ktorých staviame naše konanie, spoznávanie sveta, vedecko-technologickú činnosť i celkové fungovanie v reálnom svete. V súčasnosti moderné systémy strojového učenia na výzvu – ukáž, ako si to spravil, resp. predveď spôsob, ako pracuješ – nevedia jednoducho odpovedať (lepšie povedané, ani autori týchto systémov nevedia vo všeobecnosti objasniť, ako presne ich hlboké siete prichádzajú k výsledkom).[171]

Uvedený problém sa stáva ešte vypuklejším v rámci debát o implementácii bezpečnostných mechanizmov a etických mantinelov do systémov AI. **V tejto rovine sa sofistikovaný systém umelej inteligencie javí ako *black box* – čierna skrinka, ktorá niečo vykonáva, ale ako a prečo tak robí, nie je jasné**.

V tejto súvislosti môžeme spomenúť ešte jeden aspekt strojového učenia – tzv. hyper

---

170 Ide o vážny problém, preto existujú viaceré odborné aktivity snažiace sa black box otvoriť a jeho obsah vzniesť na svetlo, t.j. chápať, ako systém AI pracuje. Úspechy sú však len parciálne a v úzko špecifických nasadeniach algoritmov umelej inteligencie.
Napr. systém AI ExoMiner vyvinutý v NASA: „Unlike other exoplanet-detecting machine learning programs, ExoMiner isn't a black box – there is no mystery as to why it decides something is a planet or not," said Jon Jenkins, exoplanet scientist at NASA's Ames Research Center in California's Silicon Valley. „We can easily explain which features in the data lead ExoMiner to reject or confirm a planet."
New Deep Learning Method Adds 301 Planets to Kepler's Total Count. [on-line]. [cit. 28. novembra 2022].
Dostupné na internete: <https://www.jpl.nasa.gov/news/new-deep-learning-method-adds-301-planets-to-keplers-total-count>

171 Odborný časopis Technology Review z MIT (Massachusetts Institute of Technology) v tejto súvislosti uvádza, že ide o „temné tajomstvo v srdci umelej inteligencie" a dodáva: „nikto v skutočnosti nevie, ako najpokročilejšie algoritmy robia to, čo robia. To by mohol byť problém".
KNIGHT, W. *The Dark Secret at the Heart of AI*. In: *Technology Review.* [on-line]. 2017, 11. 4. [cit. 10. februára 2022].
Dostupné na internete: <https://www.technologyreview.com/2017/04/11/5113/the-dark-secret-at-the-heart-of-ai/>

parametre. Súčasné systémy umelej inteligencie prevažne pracujú na základe strojového učenia, ktoré sme vo všeobecnosti definovali ako proces učenia sa (resp. realizácie systémov AI na základe automatického/samostatného učenia sa) z existujúcich vzoriek, dát, prípadne skúseností.[172] Pre úplnosť však treba dodať, že aby sa systémy AI dokázali pomocou algoritmov strojového učenia niečo naučiť (vytrénovať/nastaviť váhy spojení neurónovej siete), treba ich najprv nadizajnovať, resp. prednastaviť pomocou tzv. hyper parametrov. Termín hyper parametre (hyperparameters) je zastrešujúcim vyjadrením pre celú množinu parametrov, ktoré musia byť človekom prednastavené, aby neurónová sieť bola vôbec schopná úspešne sa učiť. Do tejto množiny patrí napríklad počet vrstiev neurónovej siete, veľkosť recepčných polí jednotlivých uzlov konvolučných sietí, parametre rýchlosti učenia sa, použitá aktivačná funkcia a spôsob klasifikácie i veľa ďalších technických detailov.[173]

Takže v rámci voľby použitých algoritmov a rozhodnutia o komplexnom dizajne neurónovej siete je treba vyladiť (používa sa termín tuning) veľa hyper parametrov, ktorých vzájomná interakcia vedie k správnemu fungovaniu systému AI. Naviac, vo väčšine prípadov ide o unikátne „namiešanú zmes“ techník a parametrov, ktorú treba „namiešať“ nanovo pre každé nové zadanie, na ktoré bude neurónová sieť trénovaná. **Vyladenie hyper parametrov sa tak stáva absolútne kľúčovým pre správnu funkčnosť systémov AI**.[174]

V súčasnosti neexistuje univerzálny návod na vyladenie hyper parametrov v rámci návrhov systémov strojového učenia. Ide o schopnosť, ktorá je ovocím dôkladných znalostí, vytrvalej učenlivosti a mnohokrát ťažko nadobudnutej skúsenosti.[175] Eric Horvitz, riaditeľ výskumných laboratórií Microsoftu, preto na margo ladenia hyper parametrov a úspešného návrhu systému AI hovorí: „**To, čo v súčasnosti robíme, nie je veda, ale určitý druh**

172 Napr. využitie back-propagation algoritmu v systémoch učenia sa s učiteľom (supervised learning) alebo skúseností, resp. interakcie s prostredím v učení formou odmeňovania (reinforcement learning). Základné rozdelenie algoritmov strojového učenia sme popísali v kapitole 1.6.

173 Por. MITCHELL, *Artificial Intelligence*, s. 97.

174 Por. MITCHELL, *Artificial Intelligence*, s. 98.

175 Z takto namiešanej zmesi sa (nielen) v našej praxi kybernetickej bezpečnosti občas rodia konkrétne zásahy na zastavenie prebiehajúcich hackerských útokov, ktoré je niekedy ťažko explicitne popísať. Žargónom IT je to len magic – kúzlo, čím sa vyjadruje nie presný postup riešenia, ale skôr odborná intuícia, ktorá viedla k očakávanému výsledku.

alchýmie.“[176] Čo sa týka elitného klubu odborníkov, ktorí toto ladenie dokážu realizovať, Demis Hassabis, spoluzakladateľ Google DeepMind, poznamenáva: „Je to takmer ako umenie, ako z týchto systémov dostať to najlepšie... Na svete je len niekoľko stoviek ľudí, ktorí to dokážu robiť naozaj dobre.“[177]

Nech už hovoríme o čiernej skrinke, alchýmii alebo umení, **súčasné systémy AI sprevádzajú oprávnené a vážne obavy z toho, že ak nechápeme ako tieto systémy pracujú, nemôžeme im reálne dôverovať a ťažko dokážeme predpovedať okolnosti, za ktorých tieto systémy zlyhajú**.

Toto v prísnom slova zmysle môžeme povedať aj o zlyhaniach človeka, v tomto prípade však v kontexte ľudského spoločenstva staviame na určitej *teórii mysle*, teda na modeli správania sa ľudských bytostí, ktoré je na základe fundamentálnych kognitívnych schopností (rozpoznávanie objektov, chápanie scény, porozumenie reči, atď.) štandardné a vývojom overené. Avšak pri pokročilých systémoch AI nič také, ako „teóriu mysle“ nemáme.[178]

2.1.2. Zraniteľnosti, slabiny a klamanie systémov strojového učenia

V rámci niekoľkých dekád vývoja neurónových sietí a systémov strojového učenia boli postupne identifikované viaceré rizikové faktory a zraniteľnosti systémov AI. Uveďme aspoň tie podstatné:

- malá množina trénovacích dát
- nesprávne zvolená, či nekvalitná množina trénovacích dát a predsudky
- nadmerné prispôsobovanie sa tréningovým údajom
- efekt dlhého chvosta
- klamanie hlbokých sietí a ich zraniteľnosti
- povery

Malá množina trénovacích dát (training dataset)

Úspešnosť väčšiny súčasných systémov umelej inteligencie je extrémne závislá

176 METZ, C. *A New Way for Machines to See, Taking Shape in Toronto*. New York Times, Nov. 28, 2017.

177 TANZ, J. *Soon We Won't Program Computers. We'll Train Them Like Dogs*. Wired, May 17, 2016. Ide o vyjadrenie z roku 2016 – za ten čas sa v oblasti AI veľa zmenilo, no podstata zostáva...

178Por. MITCHELL, *Artificial Intelligence*, s. 109.

na rozsiahlych[179] a kvalitných súboroch správne označených trénovacích dát, resp. tréningových iterácií.[180] Bez nich sa súčasné systémy strojového učenia nedajú vytrénovať a ich nedostatok vedie v lepšom prípade k nekvalitným, v tom horšom k nesprávnym výsledkom a fatálnym zlyhaniam.[181] A ako zistíme pri ďalších typoch zraniteľností a rizikových faktorov systémov AI, problematika trénovacích dát je oveľa širšia...

Nesprávne zvolená, či nekvalitná množina trénovacích dát a predsudky (biases)

Na základe nesprávne zvolenej alebo nekvalitnej (napr. nesprávne označkovanej) množiny trénovacích dát **sa systém AI naučí robiť chybné uzávery alebo podávať výsledky „s predsudkami“**.[182]

Ide o problém nielen technologický, ale aj sociologický – predsudky, či zaujatosť v spoločnosti vedú k voľbe nesprávnemu obsahu trénovacích dát a následne k chybným výsledkom systémov AI, pričom v mnohých prípadoch hrozí, že **systémy AI trénované na zaujatých dátach môžu tieto predsudky násobiť a spôsobiť reálne škody**.

Ďalším aspektom „zaujatých“ systémov AI je znížená úspešnosť ich činnosti na základe predsudkov, čo môže mať veľmi nepríjemné dôsledky napr. v bezpečnostných systémoch, ochrane ľudských práv, sociálnej spravodlivosti a pod.

Nesprávne fungovanie zaujatého systému AI častokrát nie je problém detekovať v rámci

---

179 I táto skutočnosť poukazuje na rozdielnosť medzi hlbokým strojovým učením a ľudským vzdelávaním sa.

180 Bavíme sa o fenoméne „big data“, čo napr. pri spracovaní obrazu konvolučnými sieťami znamená mať k dispozícii rádovo milióny správne označených trénovacích obrázkov.
Por. MITCHELL, *Artificial Intelligence*, s. 99.
„Requiring so much data is a major limitation of [deep learning] today.“
NG A. *Deep Learning in Practice: Speech Recognition and Beyond*. [on-line]. In: *EmTech Digital.* 2016, 23. máj. [cit. 3. februára 2022].
Dostupné na internete: <https://events.technologyreview.com/video/watch/andrew-ng-deep-learning/>

181 Viac o fenoméne „big data“ a veľkej množine trénovacích dát je uvedené v kapitole 2.5.

182 Viac krát sa stalo, že systémy AI museli byť vypnuté, lebo po uvedení do prevádzky sa na základe nesprávne nastavených hyper parametrov a zvolenej množiny trénovacích dát vyvíjali nesprávnym smerom. Napr.:
KRAFT, *Microsoft shuts down AI chatbot after it turned into a Nazi,* [on-line]. [cit. 6. augusta 2020].
Dostupné na internete: <https://www.cbsnews.com/news/microsoft-shuts-down-ai-chatbot-after-it-turned-into-racist-nazi/>
HAMILTON, *Amazon built an AI tool to hire people but had to shut it down because it was discriminating against women,* [on-line]. [cit. 6. augusta 2020].

ostrého nasadenia, no nie je triviálne to odhaliť v predstihu (napr. v procese učenia).

V niektorých oblastiach nasadenia umelej inteligencie nie je jednoduché natrénovať systém AI bez predsudkov. Vyžaduje si to mnohokrát veľkú erudovanosť a nasadenie tých, ktorí pripravujú trénovacie dáta, pričom aj pre nich platí, že pokiaľ sú súčasťou spoločnosti akceptujúcej predsudky, podvedome ich môžu prenášať aj do svojej práce.[183] A tu sa dostávame na veľmi tenký ľad, **pretože v mnohých oblastiach sa ako spoločnosť nezhodneme na tom, čo je a čo nie je predsudok** (napr. pohlavie vs. rod s útokmi na tzv. rodové stereotypy a pod.).

Niektoré problémy s predsudkami takmer určite v budúcnosti dokážu vyriešiť systémy AGI (ak budú existovať), ktoré by mali chápať základné koncepty a v rámci abstrakcie eliminovať celú škálu zaujatostí v oblasti strojového učenia. Zvyšok, možno ten podstatný, bude výzvou rovnako, ako boj s predsudkami v spoločnosti.

Nadmerné prispôsobovanie sa tréningovým údajom (overfitting to training data)

Ide o problém, v rámci ktorého **sa systém naučí z tréningových dát rozlišovať niečo iné, než to, čo sa mal naučiť**. Napr. ak sa má systém naučiť rozlišovať nakreslený kruh a v rámci učiaceho procesu trénovacie dáta budú obsahovať vzorky, na ktorých je nakreslený kruh vždy len na zelenom podklade, je možné, že systém sa naučí ako kruh identifikovať nie to, čo nakreslený kruh skutočne obsahuje, ale skôr všetko to, čo je zelené. Ak potom pri ostrej prevádzke bude na vstupe napríklad trojuholník na zelenom podklade, systém ho identifikuje ako kruh.

V uvedenom prípade je odhalenie problému veľmi ľahké a náprava jednoduchá. Avšak v reálnom svete (napr. pri detekčných systémoch AI v kvantovej fyzike, onkológii, atď.) to vôbec nemusí byť jednoduché a korelácia, koherentnosť i kauzalita v rámci tréningového procesu vôbec nemusí byť zjavná.

Ešte jedna poznámka k doteraz uvedeným problémom s trénovacími dátami – minimálne v niektorých oblastiach pokročilých systémov sa na dôvažok javí, že ani nadmieru veľká a kvalitná množina v súčasnosti používaných trénovacích dát nebude dostatočná pre pokrok v rozvoji týchto systémov a k ich priblíženiu sa k silnej a všeobecnej AGI. Napríklad v oblasti počítačového videnia prechod od kvalitnej detekcie a rozpoznania objektov k chápaniu scény nie je možný bez detekcie a vytvárania vzťahov medzi snímanými 3D objektami, čo však nie je realizovateľné bez dostatočnej množiny

183 Por. MITCHELL, *Artificial Intelligence*, s. 106-108.

situačných videí danej scény a inovatívnych algoritmov, ktoré ich budú spracúvať, aby následne systém AI konkrétny rozpoznaný objekt (napr. mačku vo dverách, človeka na prechode) vnímal a „chápal“ v príbehu celej snímanej scény.[184]

Efekt dlhého chvosta (long-tail effect)

Týmto termínom sa v oblasti umelej inteligencie rozumie veľký rozsah možných neočakávaných situácií, s ktorými by sa systém AI mohol stretnúť. Termín „long-tail“ pochádza zo štatistiky a vyjadruje určité rozdelenie pravdepodobnosti v tvare pretiahnutého chvosta, ktorý indikuje zoznam veľmi nepravdepodobných, ale možných situácií, ktoré vo výnimočných prípadoch môžu nastať. Tieto situácie nazývame aj hraničnými/okrajovými prípadmi. Efekt dlhého chvosta môžeme ľahko ilustrovať na pravdepodobnosti vybraných situácií, s ktorými sa autonómne vozidlá môžu v reálnej prevádzke stretnúť – viď obr. č. 5, ktorý poukazuje na riziko okrajových situácií, na ktoré autonómny systém v reálnej prevádzke vôbec nebude pripravený a v danej situácii môže vykonať nesprávne rozhodnutie.

V reálnom svete jednoducho nedokážeme všetko popísať a predložiť systémom strojového učenia na vytrénovanie.[185]

Ak sa nad efektom dlhého chvosta zamyslíme, vidíme, že trénovaním (učením s učiteľom, supervised learning) nie sme schopní systém AI správne naučiť zvládať všetky hraničné situácie, keďže tieto nedokážeme dostatočne zahrnúť do trénovacích dát. A ak systém AI nie sme schopní na tieto hraničné situácie pripraviť, s veľkou pravdepodobnosťou pri ich výskyte budeme čeliť neočakávaným chybám a zlyhaniam.

V kontexte zavádzania systémov umelej inteligencie do reálneho nasadenia sa tento problém v rámci konzervatívneho prístupu rieši kombináciou špeciálnych množín trénovacích dát obsahujúcich rôzne hraničné situácie a osobitne naprogramovanými obmedzeniami pre hraničné stavy, čo však neprináša dostatočnú robustnosť v reálnej

184 Por. KARPATHY, A. *Computer vision research feels a bit stagnating in a local minimum of 2D texture recognition on ImageNet*... [on-line]. Twitter. [cit. 10. februára 2022].
Dostupné na internete: <https://twitter.com/karpathy/status/1491452689825165314>

185 „We can't realistically label everything in the world and meticulously explain every last detail to the computer.“
BENGIO, Y. *Machines Dream.* In: BEYER, D. ed. *The Future of Machine Intelligence: Perspectives from Leading Practitioners*. Sebastopol, Calif.: O'Reilly Media, 14.

prevádzke. Moderný prístup zasa kombinuje algoritmy učenia s učiteľom a učenia bez učiteľa (supervised and unsupervised machine learning), t.j. natrénovanie systému na určitej množine označených dát (labeled dataset) a následné učenie sa bez učiteľa, t.j. snaha naučiť systém kategorizovať a zoskupovať vstupné dáta do celkov, pre ktoré systém AI má konkrétne riešenia a reakcie.[186]

Obr. 6. Pravdepodobnosť výskytu niektorých situácií, s ktorými sa môže autonómne vozidlo stretnúť v prevádzke

Tento prístup koketuje s myšlienkou abstrakcie a analógie, t.j. spôsobom riešenia, v ktorom človek dokáže excelovať, ale pre strojové učenie je to stále nedosiahnuteľná méta.[187]

186 Por. MITCHELL, *Artificial Intelligence*, s. 103.

187 V tomto kontexte Yann LeCun, špičkový počítačový vedec v oblasti strojového učenia a počítačového videnia uvádza: „Učenie bez učiteľa je temnou stránkou umelej inteligencie.“ (používa výraz dark matter,

Z doteraz popísaných rizikových faktorov systémov AI vyplýva, že akékoľvek priblíženie sa k všeobecnej a silnej umelej inteligencii (AGI) prekračuje možnosti súčasného strojového učenia prostredníctvom kategorizovaných a označených tréningových dát. V kontexte súčasných algoritmov strojového učenia takmer všetko učenie by muselo prebiehať bez učiteľa (unsupervised), avšak stále nemáme k dispozícii také algoritmy učenia sa bez učiteľa, ktoré by boli, resp. v budúcnosti mohli byť dostatočne úspešné. Človek, napriek tomu, že neustále robí chyby, dokáže konať skutočne inteligentne – jednoducho má to, čo chýba všetkým súčasným systémom umelej inteligencie, a to zdravý rozum (predpoklad pre všeobecnú inteligenciu[188]), ktorého súčasťou je schopnosť abstrahovať a na základe analógií a konceptov nachádzať riešenia, myslieť a riešiť na báze nesmierneho rozsahu najrozmanitejších vedomostí, poznatkov a skúseností. Ľudská bytosť využíva zdravý rozum podvedome a spontánne v akejkoľvek oblasti života. Preto **pre mnohých je dôveryhodný systém umelej inteligencie, ktorý môže úspešne fungovať v komplexnom reálnom svete, podmienený schopnosťou mať zdravý rozum tak, ako má človek**.[189] [190]

---

ktorý sa vo fyzike používa na vyjadrenie temnej hmoty [27%], ktorá spolu s temnou energiou [68%] tvorí 95% vesmíru a je pre nás stále tajomstvom, t.j. nedokážeme ju detekovať a popísať, pozorujeme však jej účinky:-)
LECUN, Y., MISRA, I. *Self-supervised learning: The dark matter of intelligence* [on-line]. [cit. 4. februára 2022].
Dostupné na internete: <https://ai.facebook.com/blog/self-supervised-learning-the-dark-matter-of-intelligence>

188 Gary Marcus, zakladateľ spoločnosti Geometric Intelligence (získanej spoločnosťou Uber) a profesor psychológie a neurónových vied na Newyorskej univerzite hovorí: „**Predpokladom všeobecnej inteligencie je zdravý rozum; kým ho nedosiahneme, zostane nám úzka umelá inteligencia, ktorá je zriedkavo robustná a nikdy nie tak flexibilná ako ľudský rozum**."
CHADHA, S. *“Common Sense” is the Dark Matter of Artificial Intelligence*. [on-line]. [cit. 4. februára 2022].
Dostupné na internete: <https://hackernoon.com/the-dark-matter-of-ai-common-sense-is-not-so-common>

189 Por. MITCHELL,*Artificial Intelligence*, s. 104.

190 Akonáhle rozšírime svoj pohľad od učenia sa bez učiteľa k požiadavke „zdravého rozumu“, môžeme LeCunovo vyjadrenie rozšíriť: „Zdravý rozum je temnou stránkou umelej inteligencie.“
CHADHA, *“Common Sense” is the Dark Matter of Artificial Intelligence,* [on-line]. [cit. 4. februára 2022].
Dostupné na internete: <https://hackernoon.com/the-dark-matter-of-ai-common-sense-is-not-so-common>

Klamanie hlbokých sietí a ich zraniteľnosti (fooling deep neural networks and vulnerability to hacking)

Čo má spoločné špeciálny make up, pokreslená cesta, samolepkami oblepený stĺp, či avantgardná potlač na tričku? Dokážu dokonale zmiasť rôzne moderné systémy strojového videnia, detekcie objektov, rozpoznávania tvárí, či autonómnych systémov riadenia vozidiel. Žiaľ, osobitne v poslednej dekáde akcelerovaného vývoja systémov strojového učenia zisťujeme, že **je neuveriteľnej jednoduché priebežne a mnohými spôsobmi oklamať hlboké neurónové siete**.[191] Naviac, oklamanie systémov AI je možné vykonať nielen pre človeka ľahko viditeľnými a rozlíšiteľnými spôsobmi, ale častokrát i technikami, ktoré sú pre ľudskú bytosť nepostrehnuteľné. Obr. č. 7 napríklad ukazuje dvojice príkladov, na ktorých konvolučná neurónová sieť AlexNet, ktorá bola najlepšou sieťou na klasifikáciu obrázkov z databázy ImageNet v roku 2012, totálne pohorela. Nesprávne klasifikovaný obrázok z dvojice obsahuje voľným okom prakticky nepostrehnuteľné úpravy na úrovni pixelov, čo stačí na úplné pomýlenie – a ako vidíme na obrázku – aj možné cielené pomýlenie systému AI konkrétnym smerom. Pritom človek s klasifikáciou všetkých uvedených obrázkov nemá najmenší problém.

Obr. 7. Správne a nesprávne klasifikované obrázky sieťou AlexNet.[192]

Neurónové siete môžu byť však klamané aj inými spôsobmi. Napríklad – ak sa budeme

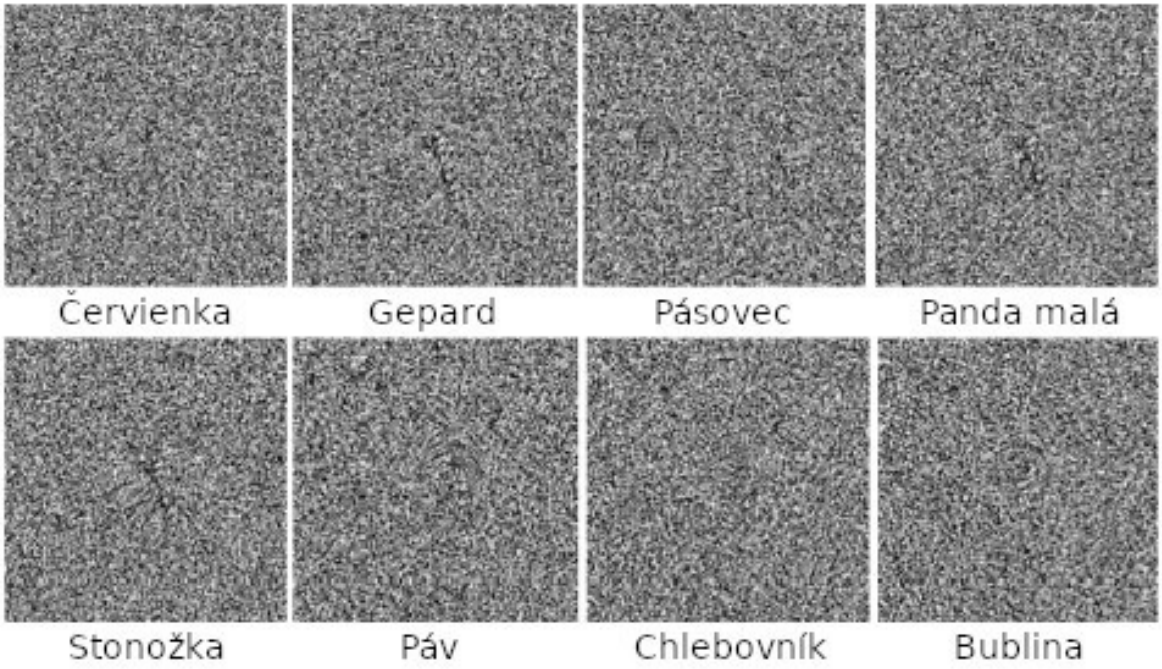
191 Por. MITCHELL, *Artificial Intelligence*, s. 110.

192 MITCHELL, *Artificial Intelligence*, s. 110, upravené autorom.

pohybovať stále v oblasti strojového videnia – pre človeka nedefinovateľné obrázky náhodného šumu môže hlboká sieť vyhodnotiť s veľkou mierou istoty ako konkrétny objekt.

Obrázok č. 8 zobrazuje príklady, ktoré vyzerajú ako náhodný šum, no pre AlexNet a iné konvolučné siete ide s viac než 99% pravdepodobnosťou o konkrétne kategórie objektov.

A aby toho nebolo málo, príklady šumu z obr. č. 8, ktoré zmiatli konvolučné siete, boli vygenerované pomocou výpočtových postupov, ktoré sa nazývajú genetické algoritmy a sú inšpirované procesmi z biologických organizmov.[193] Ide teda o obrázky, ktoré sa pomocou genetických algoritmov „vyvinuli“[194] do podoby nešpecifického šumu pre človeka, avšak pre systémy AI do podoby jasne identifikovaných kategórií objektov.



Obr. 8. Príklady šumu, ktoré konvolučné siete vyhodnocujú ako kategórie objektov.[195]

Rôzne výskumy v tejto oblasti len potvrdili, že v rámci CNN je na zlyhanie náchylná nielen AlexNet, ale zraniteľné sú aj viaceré iné konvolučné siete, a to napriek tomu, že mali

193 Por. NGUYEN, A., YOSINSKI, J., CLUNE, J. *Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images*. [on-line]. CVPR, 2015. [cit. 12. februára 2022]. Dostupné na internete: <https:// cv-foundation.org/openaccess/con tent_cvpr_2015/papers/Nguyen_Deep_Neural_Networks_2015_CVPR_paper.pdf>

194 MITCHELL, M. *An Introduction to Genetic Algorithms*. Cambridge, Mas.: MIT Press, 1996.

195 NGUYEN, YOSINSKI, CLUNE, *Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images*, upravené autorom.

rozličné architektúry, hyperparametre a množiny trénovacích dát.[196] I keď to Christian Szegedy a jeho kolegovia vo svojom vedeckom článku nazvali jednou zo „zaujímavých vlastností“ neurónových sietí, ide o reálny problém – jednoducho, konvolučné **neurónové siete sú náchylné na zlyhanie pri záškodníckych dátach** (adversarial examples).

Problém je však širší a netýka sa „len“ CNN, t.j. konvolučných neurónových sietí. Ako potvrdili ďalšie výskumy, pomocou (nielen) genetických algoritmov je možné pripraviť také vstupné dáta a cesty, ktoré podobným spôsobom oklamú hlboké neurónové siete vo všeobecnosti.[197] Tieto útoky v súčasnosti dokážu oklamať systémy identifikácie osôb, autonómnych vozidiel, spracovania lekárskych dát, rozpoznávania reči, analýzy textu a pod.

Je vážnym zistením, že **mnohé z možných útokov sú prekvapivo robustné** – dokážu účinne oklamať rôzne a diametrálne odlišné sofistikované systémy strojového učenia.[198]

Zistené zraniteľnosti pomocou záškodníckych dát tak prinášajú dilemu a poznatok.

Dilemu medzi evidentným úspechom systémov hlbokého učenia v rozličných zadaniach umelej inteligencie a jednoduchosťou, s akou je možné tieto systémy oklamať.

Poznatok, že **ani o súčasných pokročilých systémoch strojového učenia nemôžeme povedať, že ich učiaci proces je podobný tomu ľudskému a ich schopnosti nie je možné porovnávať s ľudskými, tobôž hovoriť o ich rovnocennosti alebo prekročení tých ľudských.[199] Znovu tak vyvstáva otázka – čo sa v skutočnosti tieto neurónové siete naučili? Ide skutočne o zručnosti korešpondujúce s konceptami, ktoré sa ich snažíme naučiť? Lebo napríklad pri vizuálnych systémoch je esenciálny rozdiel medzi rozpoznávaním objektov a pochopením zobrazovanej scény**. A ďalej, **chápanie nielen v kontexte scény, ale vo všeobecnosti (ako sa daný objekt správa v reálnom svete, aké má vlastnosti a pod.) sa tak javí ako nutná podmienka robustného**

196 SZEGEDY, CH. et al. *Intriguing Properties of Neural Networks*. Proceedings of the International Conference on Learning Representatins, 2014.

197 Por. NGUYEN, YOSINSKI, CLUNE, *Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images,* [on-line]. [cit. 12. februára 2022]. Dostupné na internete: <https:// cv-foundation.org/openaccess/con tent_cvpr_2015/papers/Nguyen_Deep_Neural_Networks_2015_CVPR_paper.pdf>

198 Por. MITCHELL, *Artificial Intelligence*, s. 113.

199 Súčasné najlepšie systémy CNN síce dosahujú úspešnosť v klasifikácii objektov vyššiu ako človek, ale ako sme uviedli, mýlia sa v situáciách, ktoré ľudská bytosť s prehľadom zvláda.

a spoľahlivého systému umelej inteligencie.[200] [201]

Z pohľadu zraniteľnosti systémov umelej inteligencie je problém oklamania hlbokých sietí samozrejme riešený, keďže sa vytvárajú tzv. adversarial learning algoritmy a stratégie, ktorých cieľom je ochraňovať systémy AI pred týmto typom zraniteľnosti. Je však pravdou, že **i keď sa pochopeniu a obrane voči rôznym potencionálnym útokom venuje veľa z vývoja súčasných systémov AI, na rozdiel od špecifických obranných riešení neexistuje žiadna účinná všeobecná obranná stratégia. Preto v rámci súčasnej úrovne vývoja hlbokých sietí môžeme povedať, že podobne – ako v oblasti klasickej kybernetickej bezpečnosti – ide o nekončiaci zápas medzi hľadaním nových útočných postupov a tvorbou obranných stratégií**.[202]

Z pohľadu ďalšieho rozvoja systémov umelej inteligencie ide o jasné indikovanie rozdielov medzi schopnosťami súčasných systémov ANI a riešeniami, ktoré môžu viesť k AGI, t.j. k skutočnej inteligencii ako takej.

Povery (superstition) – prirýchly Q-learning

Poverou zvykneme nazývať mylnú vieru, že určitá akcia, či úkon môžu pomôcť zapríčiniť dobrý alebo zlý výsledok. V oblasti umelej inteligencie ide o problém prevažne v rámci algoritmov učenia formou odmeňovania (reinforcement learning), pri ktorých sa trénovanie modelu (agenta) realizuje prostredníctvom interakcie s prostredím metódou pokus-omyl.[203]

V rámci trénovania systému AI (napr. robotického systému) vzniká povera vtedy, ak sa daný systém chybne naučí vykonávať nejaký nepotrebný, ba až možno nebezpečný úkon pre dosiahnutie požadovaného cieľa. Môže tak ísť napríklad o prekvapivý a nebezpečný pohyb robotického ramena alebo nečakaný manéver autonómneho vozidla, ktorý ohrozí účastníkov premávky.

Nielen vyvarovanie sa poverám, ale aj celkový návrh úspešného systému učenia formou odmeňovania je stále ešte určitou alchýmiou či umením, ktoré zvláda relatívne malá skupina expertov s veľkým citom a praxou v ladení hyperparametrov.[204] [205]

200 Por. MITCHELL, *Artificial Intelligence*, s. 112-114.

201 Znovu sa tak dostávame k problematike kognitívneho vnímania a „zdravého rozumu“.

202 Por. MITCHELL, *Artificial Intelligence*, s. 113-114.

203 O učení formou odmeňovania pojednáva kapitola 1.6.3.

204 O ladení hyperparametrov, obraznej alchýmii a umení bola rozprava v kapitole 2.1.1.

205 Por. MITCHELL, *Artificial Intelligence,* s. 138-143.

Zámerne ako poslednú z konkrétnych zraniteľností uvádzame zraniteľnosť algoritmov učenia formou odmeňovania (reinforcement learning), keďže ide o oblasť, ktorá je pokladaná asi za najperspektívnejšiu v snahe o dosiahnutie všeobecnej umelej inteligencie (AGI).

I napriek neodškriepiteľnému úspechu vývoja a nasadenia súčasných systémov umelej inteligencie v širokom spektre akademického i reálneho prostredia musíme mať neustále na pamäti, že tieto systémy môžu zlyhávať najrozličnejšími a často neočakávanými spôsobmi v dôsledku nemožnosti pripraviť dostatočne veľkú množinu trénovacích dát, prípadne ich nesprávnej voľby bez dostatočnej kvality alebo s predsudkami, nadmernému prispôsobovaniu sa tréningovým údajom, efektu dlhého chvosta, rizikám plynúcim z oklamania hlbokých sietí, ich zraniteľností a povier, pod čo sa podpisuje i nedostatok odbornej erudovanosti potrebnej pre dizajn a ladenie hyperparametrov pri príprave funkčného a úspešného riešenia.[206]

Pri hlbšom pohľade na uvedené problémy nás môžu dobiehať aj ich ďalšie dôsledky – nielen riziká priameho zlyhania, ale aj **realita výsledkov, ktoré môže byť ťažké správne interpretovať** (čo sa vlastne sieť naučila, čo výstup z daných dát na vstupe vlastne znamená) a **neschopnosť predvídať, kedy sa jednotlivé zlyhania prejavia** (za akých podmienok, pri akej súhre okolností, s dôsledku akej dynamiky vnútorného vývoja, resp. činnosti systému AI).

V zásade možno povedať, že súčasné systémy strojového učenia budú pri nasadení v reálnom svete stále trpieť neduhmi, ktoré bez dosiahnutia silnej AI nebude možné odstrániť. Totižto **úzke zameranie súčasných systémov strojového učenia je v kontraste s procesom učenia sa človeka, ktoré je vecou všetkých rozmerov ľudskej inteligencie a aktívneho zaradenia sa do sveta.**[207] **Pri zavádzaní súčasných**

206 Okrem podstatných zraniteľností a slabín uvedených v tejto kapitole technológie AI zápasia so širokým spektrom ďalších problémov. Len v rámci systémov s učením formou odmeňovania možno spomenúť také problémy ako Safe exploration, Robustness to distributional shift, Avoiding negative side effects, Avoiding “reward hacking” and “wireheading”, Scalable oversight a pod.
Por. AMODEI, D., OLAH, CH., STEINHARDT, J. et al. *Concrete Problems in AI Safety*. [on-line]. [cit. 7. júla 2022].
Dostupné na internete: <https://arxiv.org/abs/1606.06565>

207 Por. MITCHELL, *Artificial Intelligence*, s. 97.

systémov AI do nasadenia v reálnom svete si treba uvedomiť, že ich spoľahlivosť je limitovaná a schopnosti sú obmedzené. Analogicky treba s týmto rizikom narábať aj v prípade etických noriem a mantinelov, ktoré dokážeme v rámci systémov umelej inteligencie implementovať.

2.2. Umelá inteligencia v reálnom nasadení – bezpečnosť procesov

Ako sme už viackrát uviedli, systémy umelej inteligencie sa stávajú neoddeliteľnou súčasťou fungovania rodiacej sa informačnej spoločnosti a v mnohých oblastiach života sú nasadené a pomerne úspešne využívané. Druhý pohľad na limity a riziká súčasných systémov AI preto zameriame na bezpečnostné aspekty ich nasadenia a útoky, ktorým čelia.

V zásade rozlišujeme tri druhy útokov na systémy AI používané v reálnej prevádzke:[208]

- útoky na dôvernosť (confidentiality attacks)
- útoky na zraniteľnosti (evasion attacks)
- útoky s cieľom ovplyvniť model (poisoning attacks)

Útoky na dôvernosť (confidentiality attacks)

Ide o útoky zamerané na dáta uložené v rámci modelov systémov AI, ktorých zámerom je snaha kopírovať model za účelom extrahovania tréningových dát a parametrov z týchto modelov.[209] Ako sme uviedli v kapitole 2.1.2., kvalitné a rozsiahle tréningové dáta sú jedným z podstatných faktorov správne fungujúceho systému AI. Pre reálne nasadenie v konkrétnom sektore, resp. organizácii sa tak súčasťou tréningových dát môžu stať dôverné, osobné, prípadne strategické informácie.

Naviac v kontexte vysokej sofistikovanosti kvalitných systémov strojového učenia je nezanedbateľným dôvodom aj snaha získať informácie o dizajne a hyper parametroch konkrétneho systému AI.

Útoky na zraniteľnosti (evasion attacks)

208 REHÁK, M. *Útoky na systémy umělé inteligence a jejich obrana*. In: *Umělá inteligence 2021*. Praha: TUESDAY Business Network, 2021.

209 Jedným z jednoduchých spôsobov útoku je napr. cielené zadávanie veľkého kvanta vstupných dát do systému konkurencie a sledovanie/analýza výsledkov.
REHÁK, *Útoky na systémy umělé inteligence a jejich obrana*.

V tejto oblasti sa útočníci zameriavajú na odhaľovanie a zneužitie existujúcich zraniteľností v modeloch za účelom zmanipulovania výsledkov systémov AI. Ide tak – na základe existujúcich zraniteľností a limitov – o ovplyvňovanie činnosti systémov AI priamo počas ich prevádzky.[210]

Útoky s cieľom ovplyvniť model (poisoning attacks)

Ide o celú škálu útokov, ktorých cieľom je ovplyvňovanie modelu, tréningového procesu a tým aj výslednej činnosti systému AI. Ide o zámerné ovplyvňovanie tréningového procesu (učenia) modelu za účelom manipulovania jeho následných rozhodnutí v prevádzke.

V tejto súvislosti je zaujímavým etickým problémom, ktorý sme doteraz nespomínali, aj iný rozmer zámerného ovplyvňovania systému AI – zámerné nastavenie parametrov a ovplyvňovanie tréningového procesu samotnými tvorcami daného systému, či už zo svojej vôle alebo na základe zadania objednávateľa systému. Osobitne v prípade, keď ide o vládny subjekt, nadnárodnú spoločnosť, celosvetovú sociálnu sieť, informačné platformy a pod., sa jedná o vážny a delikátny problém. Avšak problém, ktorý je – žiaľ – reálny.[211]

Ak sa pozrieme na pomerné zastúpenie výskytu jednotlivých druhov útokov, prax poukazuje na veľkú rôznorodosť výskytu a použitia.

Útokov na dôvernosť (confidentiality attacks) je v súčasnosti minimum, pretože ich cieľ sa dá dosiahnuť inými spôsobmi, ktoré sú mnohokrát jednoduchšie, časovo menej náročné a lacnejšie. Keďže väčšina organizácií útoky tohto druhu neočakáva, ich nasadenie môže byť oveľa väčším rizikom s nedobrým koncom. Preto treba byť opatrný.

Úspešnosť útokov s cieľom ovplyvniť model (poisoning attacks) sa v súčasnosti viaže

210 Oklamanie systému na rozpoznávanie tvárí, podvodné získanie financií, resp. úveru z bankového domu, vyradenie z činnosti systému pre automatické riadenie streľby alebo skupinového riadenia dronov, atď.

211 V súčasnosti ide prevažne o problém, s ktorým sa potýkajú systémy umelej inteligencie nasadené v oblasti spracovania informácií a informačných tokov.
Viac o možnostiach zneužitia informácií napr. v kapitole *Informácia ako nástroj moci* práce:
ŠANTAVÝ, *Analýza virtuálneho sveta v kontexte náboženskej formácie, pôsobenia a života Cirkvi,* [on-line], s. 37-39. [cit. 14. februára 2022].
Dostupné na internete:
<https://peter.santavy.cloud/data/uploads/docs/analyza_virtualneho_sveta_v_kontexte_nabozenskej_formacie_posobenia_a_zivota_cirkvi.pdf>

na pokročilé znalosti z oblasti umelej inteligencie, preto sa tento druh skôr nahrádza jednoduchšími útokmi zacielenými na zraniteľnosti systémov AI.

Potenciál útokov s cieľom ovplyvniť model je však veľký a osobitne s možnosťou nástupu všeobecnej a silnej umelej inteligencie ich atraktívnosť a nutkanie ich zneužiť môže veľmi rásť.

Útoky na zraniteľnosti (evasion attacks) tvoria v súčasnosti takmer 100% útokov. Vzhľadom na neustály vývoj a otvorenú komunikáciu ohľadom zraniteľností systémov AI býva mnohokrát dostatok informácií pre zneužitie konkrétnych zraniteľností (vytvorenie exploitu) a realizácia útokov je pomerne jednoduchá, takže sa dajú ľahko vykonať.[212]

V kontexte bezpečnosti procesov umelej inteligencie sa etické výzvy, doteraz rozoberané ako súčasť návrhu a realizácie systémov AI, rozširujú aj o oblasť etiky použitia, resp. riziká zneužitia. V oblasti klasickej informačnej bezpečnosti sú na jednej strane barikády tí, ktorí bezpečnosť systémov strážia (profesionáli v oblasti kybernetickej bezpečnosti), odhaľujú zraniteľnosti s cieľom nápravy a ochrany systémov (napr. etickí hackeri), riešia škody a ochraňujú obete, na strane druhej je skrytý zástup od jednotlivcov až po organizované skupiny, ktorí sa snažia systémy zneužiť, znefunkčniť, či vykradnúť dôležité dáta z veľmi rôznorodých dôvodov (od ideologických, psychologických, sociologických až po mocenské dôvody a kybernetickú kriminalitu[213]). Nie inak je tomu tak aj v oblasti umelej inteligencie, keď prekročenie etického rámca je pre mnohých príliš

212 To neznamená, že by bolo vhodnejšie hľadanie, analýzu a riešenie zraniteľností skrývať. Už v rámci klasickej kybernetickej bezpečnosti je overeným faktom, že bezpečnosť prostredníctvom utajenia (security by obscurity) je skutočne zlý nápad. Pri otvorenom prístupe k problému zraniteľností sa tieto môžu rýchlejšie identifikovať, komplexnejšie analyzovať a plošne odstraňovať.

213 Informačná kriminalita v súčasnosti patrí spolu s drogami a obchodovaním s ľuďmi medzi tri najvýnosnejšie oblasti kriminality.
ŠANTAVÝ, *Analýza virtuálneho sveta v kontexte náboženskej formácie, pôsobenia a života Cirkvi,* [on-line], s. 37. [cit. 14. februára 2022].
Dostupné na internete:
<https://peter.santavy.cloud/data/uploads/docs/analyza_virtualneho_sveta_v_kontexte_nabozenskej_formacie_posobenia_a_zivota_cirkvi.pdf>
Ba čo viac – už v rokoch 2015/16 sa kybernetická kriminalita stala oblasťou najvýnosnejšou.
KHIMJI, I. *Cybercrime Is Now More Profitable Than The Drug Trade* [on-line]. [cit. 14. februára 2022].
Dostupné na internete: <https://www.tripwire.com/state-of-security/regulatory-compliance/pci/cybercrime-is-now-more-profitable-than-the-drug-trade/>

lákavé pokušenie, ktorému nebudú vedieť a – ak sa etika návrhu a využívania systémov AI nebude zodpovedne riešiť – ani nebudú chcieť a mať prečo odolať.

A keď sme už pri etike, na verejnosť neustále presakujúce informácie o únikoch z vládnych zdrojov a bezpečnostných agentúr dávajú tušiť, že pokušenie je priveľké aj pre tak veľkých „hráčov“ v oblasti nasadzovania systémov AI, ako sú vlády, nadnárodné spoločnosti a mocenské zoskupenia, keď zápasia (*sic!*) s túžbou využitia (zneužitia) umelej inteligencie ako nástroja na dosiahnutie neprimeraného zisku, ideologických a mocenských cieľov.[214]

2.3. Kybernetická bezpečnosť máta aj umelú inteligenciu

Súčasné systémy umelej inteligencie sú postavené na moderných informačných a komunikačných technológiách. Preto – ako elektronické systémy – zdieľajú aj všetky riziká a neduhy moderných kybernetických systémov.[215] Tretí pohľad na limity a riziká súčasných systémov AI sa preto zameriava na kybernetickú bezpečnosť riešení postavených na umelej inteligencii.

Naviac, **kybernetická bezpečnosť je oblasťou, ktorá určitým spôsobom doteraz uvedené problémy prepája, keďže v reálnom nasadení vo väčšine prípadov nie je možné riziká striktne rozdeľovať podľa vyššie uvedených kapitol. Mnohé riziká napríklad atakujú bezpečnosť procesov umelej inteligencie, zneužívajú nedostatky dizajnu alebo amplifikujú dôsledky rizikových faktorov, pričom však ako vektor útoku používajú zraniteľnosti v oblasti kybernetickej bezpečnosti.**

Napríklad jedna z najpoužívanejších platforiem pre vývoj a prevádzku systémov strojového učenia, TensorFlow od Googlu mala len za rok 2021 oficiálne evidovaných 201 zraniteľností (pridelené CVE[216]). Je zaujímavé, že z pohľadu kybernetickej bezpečnosti má

214 Viac o týchto problémoch pojednávame od kapitoly 2.5 ďalej.

215 Uvádzame konkrétnejší pojem kybernetický systém, v ktorom sa snúbi skĺbenie výpočtovej techniky a kybernetického informačného priestoru (cyberspace). Pojem informačný systém je teda jeho nadmnožinou, zahŕňajúcou aj oblasti mimo počítačového sveta.

216 Systém CVE (Common Vulnerabilities and Exposures) poskytuje referenčnú metódu pre verejne známe zraniteľnosti a ohrozenia informačnej bezpečnosti. Tento systém spravuje Národné centrum kybernetickej bezpečnosti Spojených štátov amerických (NCF), ktoré prevádzkuje spoločnosť The Mitre Corporation, a financuje ho Národná divízia kybernetickej bezpečnosti amerického ministerstva pre vnútornú bezpečnosť. V oblasti kybernetickej bezpečnosti ide o relevantnú, bežne používanú

väčšina týchto zraniteľností nízke rizikové skóre (CVSS Score) – dôsledky ich zneužitia tak nemusia byť veľkým rizikom a viesť ku klasickým kybernetickým kompromitáciám (hacknutiam). Vo viacerých prípadoch môže však byť ovocím ich zneužitia prerušenie prebiehajúcej úlohy alebo jej nesprávne vykonanie. Takže nízke CVSS skóre v oblasti kybernetickej bezpečnosti nemusí znamenať malé riziko pre prevádzku systému AI.[217]

S postupným rozširovaním využívania tejto platformy a celkovo systémov AI badať aj rapídny medziročný nárast evidovaných zraniteľností,[218] čo na jednej strane znamená viac potencionálnych spôsobov kompromitácie, na strane druhej je však indikátorom aj zvýšeného záujmu o bezpečnosť softvérového vybavenia pre masívne zavádzanie AI.

Indikátorom skutočného rizika nie je tak ani počet objavených a zdokumentovaných zraniteľností, ale skôr priemerný čas potrebný na ich odstránenie.

Dôležitým faktorom je tiež spôsob nasadenia technologického vybavenia systémov AI, keďže v mnohých prípadoch ide o systémy, ktoré po nasadení do prevádzky nie sú predmetom údržby a prípadných aktualizácií (opráv) bezpečnostných chýb, takže sa stávajú bezpečnostným rizikom.[219] Neaktualizovaný jednoduchý inteligentný asistent (vznešene sa nazývajúci umelou inteligenciou vzhľadom na použité jednoduché prvky z AI) vo fotoaparáte nie je problémom, no systém, ktorý napr. riadi dopravu a nie je ťažké ho zmiasť, resp. navodiť dopravný kolaps a nehody – ak sa objavené chyby v jeho dizajne, či implementácii neošetria – môže byť veľmi nebezpečný a možné obete na životoch by boli len otázkou času, kým dané chyby niekto nezneužije, resp. sa nevyskytne súbeh

---

a celosvetovo akceptovanú metódu klasifikácie zraniteľností.
Common Vulnerabilities and Exposures. [on-line]. [cit. 15. februára 2022].
Dostupné na internete: <https://en.wikipedia.org/wiki/Common_Vulnerabilities_and_Exposures>

217 *Google » Tensorflow: Security Vulnerabilities* [on-line]. [cit. 11. januára 2022].
Dostupné na internete: <https://www.cvedetails.com/vulnerability-list/vendor_id-1224/product_id-53738/Google-Tensorflow.html>

218 *Google » Tensorflow: Vulnerability Statistics* [on-line]. [cit. 11. januára 2022].
Dostupné na internete: <https://www.cvedetails.com/product/53738/Google-Tensorflow.html?vendor_id=1224>

219 Toto je jedno z vážnych rizík nastupujúceho masívneho využívania internetu vecí (IoT), ktoré s príchodom 5G sietí smerujú k intenzívnej interkonektivite a tým aj k on-line zraniteľnosti. Zatiaľ ani neexistuje pevný rámec, ktorý by tento problém riešil, aj keď viacero iniciatív v tejto oblasti existuje. Doterajšie problémy v oblasti kybernetickej bezpečnosti s IoT poukazujú na veľké výzvy v blízkej budúcnosti.

podmienok (race condition), ktoré tento systém AI pomýlia.

Ako každá iná oblasť kybernetickej, resp. vo všeobecnosti informačnej bezpečnosti i oblasť bezpečnosti systémov AI je proces. A ako v každej inej oblasti informačných technológií, ani v oblasti systémov umelej inteligencie neexistuje dokonale bezpečný a spoľahlivý systém.

Okrem snáh o riešenie bezpečnosti systémov AI je jednou z veľkých úloh aj hľadanie spôsobov, ako minimalizovať dôsledky týchto zlyhaní či už pevnými obmedzeniami, dohľadovými riešeniami alebo inými prostriedkami. Zaujímavosťou je, že v súčasnosti takmer všetky moderné nástroje používané v rámci kybernetickej bezpečnosti na detekciu, dohľad a ochranu pred útokmi taktiež obsahujú prvky umelej inteligencie:-)

Tiež si treba uvedomiť, že počet zraniteľností je priamo úmerný komplexnosti systémov – ktorá je v oblasti umelej inteligencie jedným z normatívnych faktorov sofistikovaných a úspešných systémov. Podobne, počet kompromitácií, resp. zneužití týchto zraniteľností priamo rastie s mierou nasadenia a využívania v reálnom svete.

2.4. Technologická komplexnosť a potrebná infraštruktúra

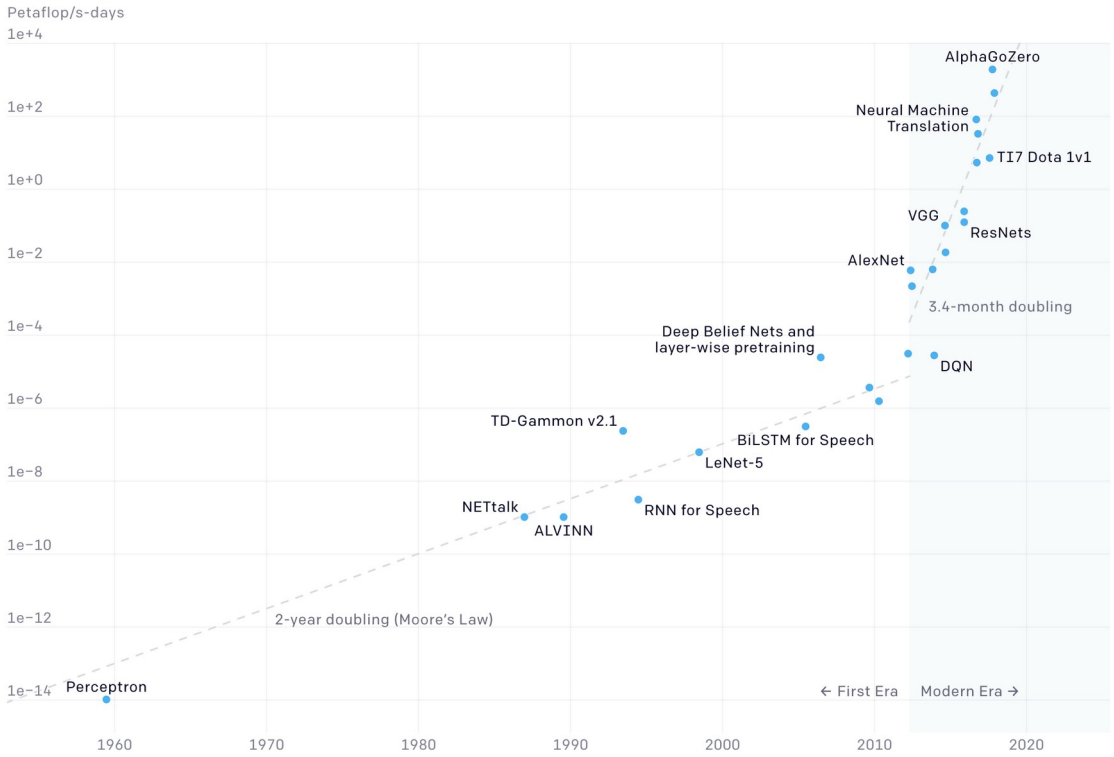
Ako sme uviedli v predchádzajúcej kapitole, súčasné systémy umelej inteligencie sú postavené na moderných informačných a komunikačných technológiách, ktoré sú tvorené nielen softvérovým vybavením (počítačové programy), ale i hardvérom, t.j. elektronickým a vo všeobecnosti technologickým zázemím, vďaka ktorému softvér systémov umelej inteligencie môže pracovať. Štvrtý pohľad na limity a riziká súčasných systémov AI sa preto zameriava na hardvérové vybavenie a technologickú infraštruktúru potrebnú pre zabezpečenie spoľahlivej funkčnosti a robustnosti systémov umelej inteligencie.

Výpočtové systémy prevádzkujúce algoritmy AI – vzhľadom na postupy strojového učenia – musia disponovať veľkým výpočtovým výkonom[220] a vzhľadom na enormné množstvo spracúvaných dát aj schopnosťou zhromažďovať, ukladať a manipulovať s veľkými štruktúrovanými i neštruktúrovanými dátami (viď príloha č. 2 – štruktúrované

220 **S rozvojom algoritmov strojového učenia sa požadovaný výpočtový výkon v poslednej dekáde zvyšuje exponenciálne!**
AMODEI, D., HERNANDEZ, D. *AI and Compute*. [on-line]. [cit. 19. januára 2022].
Dostupné na internete: <https://openai.com/blog/ai-and-compute/>

a neštruktúrované dáta).

Operácie vykonávané neurónovými sieťami tvoria pomerne špecifickú podmnožinu funkcionality, ktorú vedia moderné počítačové systémy ponúknuť, pričom ich výkon nie je špeciálne optimalizovaný primárne pre vykonávanie operácií napríklad strojového učenia. Kvôli optimalizácii výkonu sa preto v posledných rokoch vyvíja a vyrába špecializovaný hardvér[221], ktorý ponúka oveľa vyšší výkon v operáciách algoritmov umelej inteligencie než bežné – i keď výkonné – počítačové vybavenie.[222]



Obr. č. 9. Výpočtový výkon systémov AI sa v poslednej dekáde zvyšuje exponenciálne.[223]

221 Procesory optimalizované pre AI, hardvérové akcelerátory, atď.: ASIC - Application Specific Integrated Circuit, špecializované integrované obvody navrhnuté na vykonávanie úloh spojených s AI, TPU - Tensor Processing Unit, špecializovaný akcelerátor od Google, využívanie výpočtového výkonu cloudov,...

222 Trh s týmto hardvérom je pomerne rozsiahly – od superpočítačov, cez výkonné systémy autonómnych vozidiel, vojenskú techniku a rôznorodé technické vybavenie, špecializované integrované obvody (čipy) pre osobné počítače, tablety a mobily, až po napríklad špičkové a výkonné nástroje pre softvérových vývojárov a vedcov v oblasti AI, ktoré je možné kúpiť za lacný peniac (desiatky Eur) a ktoré sa dajú k počítaču jednoducho pripojiť cez USB, PCIe alebo M2 rozhranie.

223 AMODEI, HERNANDEZ, *AI and Compute*. [on-line]. [cit. 19. januára 2022]. Dostupné na internete: <https://openai.com/blog/ai-and-compute/>

Je to nutné, ak uvážime, že bez potrebného výkonu nie je možné zabezpečiť nasadenie systémov AI tak, aby pracovali v reálnom čase.

Z apetítu súčasných moderných systémov AI sa javí, že ich ďalší rozvoj – osobitne v túžbe po dosiahnutí uvedomelej AGI – bude pravdepodobne vyžadovať nielen kvantitatívnu, ale predovšetkým kvalitatívnu zmenu v schopnostiach hardvéru výpočtových systémov.[224] [225]

Technológie, potrebné pre úspešné nasadenie systémov AI, netvorí len výkonný hardvér schopný vykonávať extrémne množstvo špecializovaných operácií za sekundu. Viaceré oblasti nasadenia v reálnom svete poukazujú i na ďalšie komponenty, na ktoré nesmieme zabúdať – **špecifické periférne zariadenia a technológie vysoko rýchlostného prepojenia všetkých častí funkčného celku AI v danej oblasti nasadenia**.

Využívanie systémov umelej inteligencie v reálnom svete je pre mnohé scenáre použitia spojené s kvalitnými perifériami, t.j. vstupnými a výstupnými systémami.[226] Požadovaná kvalita a výkon periférií sa dosahuje ich neustálym zlepšovaním, implementovaním špecializovaných systémov AI (napr. na vylepšenie počítačového videnia z kamery) a orchestráciou rôznych typov periférií do spoločného celku.

Veľký objem zo senzorov generovaných a spracúvaných dát v reálnej prevádzke systémov AI a vysoké dátové toky medzi jednotlivými časťami týchto systémov, ktoré musia byť realizované v zlomkoch sekúnd, vytvárajú veľké nároky na spôsoby vysoko rýchlostného prepojenia (napríklad tzv. zbernice) v rámci jednotlivých systémov AI.[227]

---

224 Vývoj sa uberá rôznorodým smerom: či už ide o hľadanie nový polovodičových architektúr, neuromorfné čipy, optické systémy, biotronické systémy, kvantové počítače,...

225 „Kľúčovým prvkom pokroku v oblasti umelej inteligencie sú zlepšenia v oblasti výpočtovej techniky, takže pokiaľ bude tento trend pokračovať, oplatí sa pripraviť na dôsledky systémov, ktoré sú ďaleko za dnešnými možnosťami.“
AMODEI, HERNANDEZ, *AI and Compute*. [on-line]. [cit. 19. februára 2022]. Dostupné na internete: <https://openai.com/blog/ai-and-compute/>

226 Napr. autonómne vozidlá pre svoju prevádzku vyžadujú detekciu okolia a monitorovanie trasy. Na detekciu sa v súčasnosti využíva kombinácia radarov, lidarov, satelitnej navigácie, počítačového videnia a zvukový vstupov.

227 Napr. v automobilovom priemysle sa používajú uzavreté vnútrovozidlové zbernice CAN (controller area network), prípadne modernejšie FlexRay, LIN, MOST, ktoré sú však pre veľké objemy dát systémov AI v autonómnych vozidlách nedostatočné. V nich sa zavádzajú automotive ethernet, t.j. nové vysoko rýchlostné zbernice typu ethernet (dobre známe a široko využívané v počítačových sieťach).

Vo viacerých prípadoch nasadenia však musia systémy AI komunikovať aj s inými systémami, napr. autonómne vozidlá so satelitnými navigačnými systémami, s inteligentnou infraštruktúrou vozoviek, s ostatnými vozidlami, internetom, atď.[228] A to všetko v reálnom čase.[229]

Krátky náčrt problematiky technologického vybavenia a nutnej infraštruktúry pre nasadenie systémov AI vo viacerých oblastiach reálneho sveta poukazuje na veľkú komplexnosť týchto systémov. A komplexnosť je problém – je rizikovým faktorom bezpečnosti a stability fungovania systémov.

Nadviažuc na tému predchádzajúcej kapitoly o kybernetickej bezpečnosti systémov AI, v kontexte komplexnosti týchto systémov si musíme uvedomiť aj **rapídny nárast tzv. útočnej plochy (attack surface)[230] na tieto systémy**. Ide o rozšírenie možností kybernetických útokov o ďalšie typy vzhľadom na komplexnosť a množstvo použitých technológií i potrebu sofistikovanej infraštruktúry.[231]

K faktorom, ktoré asi nikdy nedokážeme naplno ošetriť, patrí aj zlyhanie jednotlivých komponentov a elektronických súčiastok. Ide o malý bonus k doteraz uvedeným problémom, ktorý sa rieši rôznymi metódami redundancie, štatistickými metódami a prvkami kontroly, pričom miera zložitosti a finančnej náročnosti pri systémoch zabezpečených voči zlyhaniam elektroniky mnohokrát neúmerne rastie. **Ekonomický rozmer tak môže veľmi ovplyvňovať aj bezpečnosť systémov AI**.

Keďže komplexnosti sa nedá vyhnúť (v zásade ide o analógiu s biologickými systémami), riešením je vývoj techník zabezpečujúcich stabilitu a bezpečnosť (napr. redundancia,

228 Ide o tzv. CCAM - coperative connected automated mobility, t.j. vzájomné prepojenie systémov účastniacich sa dopravnej prevádzky a proaktívna výmena informácií medzi nimi, čo je kľúčové pre nasadenie stupňov 4 a 5 autonómie vozidiel (viď. obr. č. 11 v nasledovnej kapitole).
Ide o tzv. V2X komunikáciu, resp. prepojenie, pričom V2X = V2N + V2V + V2I + V2P (komunikácia vehicle to vehicle, network, infrastructure, pedestrian).

229 Bezdrôtová časť komunikácie vyžaduje vysoké rýchlosti a nízke latencie, čo spĺňajú technológie VANET (Vehicular Ad Hoc Networks – Wifi medzi vozidlami na 5.9 GHz) a 5G siete. Plnú podporu by mali zabezpečiť budúce 6G siete nielen s vyššími rýchlosťami a nižšou latenciou než 5G, no predovšetkým s implementovanou sadou funkcionalít využiteľných aj pre autonómne vozidlá (tzv. AI enabled networks).

230 *Attack surface*. [on-line]. [cit. 19. februára 2022].
Dostupné na internete: <https://en.wikipedia.org/wiki/Attack_surface>

231 Napríklad rádiové rušenie alebo zahltenie (DoS/DDoS útoky), výpadok satelitnej navigácie a internetového pripojenia, špecifické útoky na internet vecí (IoT) a pod.

mnoho úrovňové bezpečnostné mechanizmy a pod.).

Čo však v rámci živých organizmov ako mechanizmy ochrany do DNA vložila evolúcia, vo svete umelej inteligencie musia v rámci technologického rozvoja zabezpečiť ľudia – tvorcovia a výrobcovia systémov. A spoľahlivú funkčnosť, ktorú u biologických systémov preverila príroda, musí v oblasti AI zabezpečiť spoločnosť svojimi pravidlami a reguláciami.

V pohľade na technologickú komplexnosť a nutnosť fungujúcej infraštruktúry (kapitola 2.4.), vzhľadom na rôznorodosť kybernetických útokov a reálnu nemožnosť vytvoriť absolútne bezpečný systém (2.3.), zároveň vo vedomí procesných útokov (2.2.) a rizikových faktorov systémov umelej inteligencie (2.1.), by mala byť **nutnou podmienkou ich prevádzky schopnosť a možnosť človeka prebrať kedykoľvek kontrolu nad týmito systémami, resp. právo a možnosť verifikovať a prehodnotiť výsledky ich činnosti.**[232]

2.5. Spoločenské dôsledky, ktoré prinášajú vrásky

Našu rozpravu o limitoch a rizikách súčasných systémov umelej inteligencie môžeme rozšíriť pohľadom na oblasť, ktorá sa prakticky dotýka každého jedného človeka modernej spoločnosti – ide o dôsledky a riziká spojené s využívaním prostriedkov AI v našom každodennom živote a interakciou, ktorú vedome či nevedome s nimi neustále podstupujeme. Ide o metamorfózy, v ktorých sa trhové mechanizmy, zábava, vzdelávanie, zdravotníctvo, doprava, vedecko-technický rozvoj, služby a spoločenské procesy, dopované umelou inteligenciou, dotýkajú nášho človečenstva.[233]

V modernej informačnej spoločnosti sa základnou komoditou stávajú informácie.[234]

232 V mnohých aplikáciách systémov AI však reálne uskutočniteľná možnosť verifikovať výsledky ich činnosti je **takmer neriešiteľný problém**.

233 V tom lepšom prípade len dotýkajú, v horšom ho možno i menia a pretvárajú.

234 Informácia ako taká sa v informačnej spoločnosti stáva *základnou komoditou jej fungovania*. Je nielen predmetom činnosti znalostnej ekonomiky, ale stáva sa i nástrojom moci, či drogou informačnej závislosti.
ŠANTAVÝ, *Analýza virtuálneho sveta v kontexte náboženskej formácie, pôsobenia a života Cirkvi,* [on-line], s. 30. [cit. 7. decembra 2021].
Dostupné na internete:
<https://peter.santavy.cloud/data/uploads/docs/analyza_virtualneho_sveta_v_kontexte_nabozenskej_formacie_posobenia_a_zivota_cirkvi.pdf>

Systémy umelej inteligencie sú priamo závislé na kvalitných informáciách, pričom o väčšine aktuálne najúspešnejších sofistikovaných systémoch AI sa dá povedať, že sú závislé na extrémnom množstve relevantných dát.

Aby nasadenie systémov umelej inteligencie bolo v spoločnosti úspešné, potrebné dáta musia byť neustále zhromažďované z reálneho sveta a priamo z ľudského prostredia.

Nutnou súčasťou extrémne rýchleho súčasného vývoja v oblasti umelej inteligencie je fungujúca možnosť monetizácie nasadenia systémov AI, čo sa najmarkantnejšie prejavuje v on-line systémoch a produktoch technologických gigantov (sociálne siete, vyhľadávače,...). U týchto systémov by sa pri povrchnom pohľade mohlo javiť, že základnou komoditou sú informácie, ktoré uvedené spoločnosti bezprecedentným spôsobom zhromažďujú, analyzujú a spracúvajú. Avšak systémy umelej inteligencie so svojimi schopnosťami posúvajú túto časť paradigmatickej zmeny informačnej spoločnosti ešte ďalej: **z informácií a z produktov ich spracovania sa komoditou stávajú priamo ľudia – ba čo viac, ich psychologické profily a vzťahy, konanie a jeho ovplyvňovanie a v konečnom dôsledku i fungovanie celej spoločnosti**.

~~Dokonale~~ (povedzme, že len tak kvalitne, ako to umožňujú viaceré súčasné platformy sociálnych sietí) natrénovaná AI sa dokáže zamerať na človeka, konkrétne sociálne skupiny, resp. vo všeobecnosti nás ľudí v spoločnosti a – analogicky využitiu kvalitných psychologických metód – priviesť nás k pozornosti, akceptácii ňou predkladaných informácií (osobitne cez pokročilé metódy, napr. augment reality, podprahové metódy,...) a k **ovplyvňovaniu nášho vnímania, našich rozhodnutí a konania**...[235] A môže to ísť až tak ďaleko, že prichádza k zmene nášho zmýšľania a vnímania, kým vlastne sme. Druhou stranou mince je **schopnosť systémov AI vytvárať modely, ktoré sú schopné predpovedať naše konanie**.[236]

Ak dosiahnutý efekt a vplyv môže byť tak závažný a prakticky takmer istý, existuje riziko

235 Por. rozhovor s Jaronom Lanierom, počítačovým vedcom a spolu zakladateľom oblasti virtuálnej reality. ORLOWSKI, J. *The Social Dilemma*. [filmový dokument]. Netflix, 2020, 14:25. [cit. 7. decembra 2021]. Dostupné na internete: <https://www.netflix.com/sk/title/81254224>

236 Por. rozhovor s Azom Raskinom, pôvodným povolaním matematikom a fyzikom, spoluzakladateľom Centra pre humánne technológie.
ORLOWSKI, *The Social Dilemma*, 17:50. [cit. 7. decembra 2021].
Dostupné na internete: <https://www.netflix.com/sk/title/81254224>

i pokušenie využiť to, manipulovať a zneužiť (či už vo svete biznisu, ale aj v oblastiach oveľa kritickejších).

Ak efekt má byť takmer istý, treba stavať na extrémne presných predikciách. A extrémne presné predikcie nie je možné vykonávať bez extrémne veľkého množstva dát[237] a techník, ktoré ich dokážu v reálnom čase spracovať.

S tým súvisí ďalšie riziko - tzv. kapitalizmus dohľadu (surveillance capitalism), ktorý ťaží z nekonečného sledovania - **ľudia i celá spoločnosť sú vystavení neustálemu dohľadu a sledovaniu, ktoré je však len veľmi málo pod kontrolou, ak vôbec**. Osobitne tak môže byť v kontexte systémov umelej inteligencie, u ktorých zhromažďované dáta nie sú takmer vôbec pod ľudským dohľadom (ak taký dohľad je pri veľkých systémoch zhromažďujúcich denno denne extrémne množstvo dát[238] vôbec možné zaručiť).

Ide o začarovaný kruh, čím viac dát systémy AI dokážu získať (a na nich sa učiť), tým presnejšie modely nášho správania ponúkajú a tým lepšie dokážu ovplyvňovať naše vnímanie, postoje a rozhodnutia – rozhodnutia i činnosť, čo následne generuje ďalšie dáta, zberané a využívané systémami AI... Na základe dát a výsledkov systémov AI sa upresňujú algoritmy implementované v týchto systémoch, aby požadované výsledky boli stále presnejšie (a modely nášho správania autentickejšie).

A riziká i možné dôsledky sú v mnohom alarmujúce...

Neuvedomujúc si, ako je naša myseľ a psychika zraniteľná, tak postupne prechádzame od technologického prostredia založeného na systémoch AI k prostrediu založenému na závislosti a manipulácii.

Potenciál a skutočnú silu týchto psychologických zásahov, vyjadruje i skúsenosť ľudí pracujúcich v oblasti implementácie systémov AI v informačnej spoločnosti, osobitne v rámci sociálnych sietí. Mnohí z nich do detailov poznajú spôsob fungovania nasadených

237 Por. rozhovor s profesorkou Shoshanou Zuboff, autorkou knihy The Age of Surveillance Capitalism.
ORLOWSKI, *The Social Dilemma*, 15:26. [cit. 7. decembra 2021].
Dostupné na internete: <https://www.netflix.com/sk/title/81254224>

238 Už v roku 2010 Erich Smid, v tom čase CEO Google, na konferencii Techonomy v kalifornskom Lake Tahoe vyhlásil: „V súčasnosti každé dva dni vyprodukujeme toľko informácií, koľko sme vytvorili od úsvitu civilizácie až do roku 2003.“
SIEGLER, *Eric Schmidt: Every 2 Days We Create As Much Information As We Did Up To 2003* [on-line]. [cit. 12. decembra 2021].
Dostupné na internete: <http://techcrunch.com/2010/08/04/schmidt-data/>

systémov AI, poniektorí z nich dokonca tieto systémy vyvíjali a nasadzovali, no sami sa stali obeťami ich fungovania. Otvorene hovoria o obdobiach svojho života, v ktorých zakúsili totálnu neschopnosť vymaniť sa z ich vplyvu a závislosti.[239]

Dr. Anna Lembke, lekárska riaditeľka liečby závislostí na Stanfordskej univerzite, poukazuje na potenciál závislosti, ktorú môžu vytvárať moderné sociálne siete, postavené na systémoch AI a dostatočne sofistikované, aby virtualizovali a nahrádzali naše reálne vzťahy: „Sociálne médiá sú droga. Máme základnú biologickú potrebu byť v kontakte s inými ľuďmi, čo má priamy vplyv na uvoľňovanie dopamínu v mezilimbickej dráhe. Tento systém, ktorý majú na svedomí milióny rokov evolúcie, sa nás snaží spájať a núti nás žiť v komunitách, aby sme si mohli nájsť partnerov a množiť sa. Takže niet pochýb o tom, že niečo ako sociálne médiá, ktoré umožňujú spojenie medzi ľuďmi, budú mať potenciál vyvolať závislosť.“[240]

So závislosťou prichádzajú aj ďalšie súvisiace problémy. Algoritmy sociálnych médií sú postavené na neustálych podnetoch, ktoré súvisia s uvoľňovaním dopamínu a pozitívnou reakciou našej mysle. Vytvára sa tak túžba po neustálom pozitívnom ohodnotení, ktoré už nie je primeranou a hlavne adekvátnou súčasťou vnemov v reálnom svete, ale umelo držanou hladinou, na ktorej sa používatelia stávajú závislí. Znižuje sa tak schopnosť prijímať a adekvátne spracovávať negatívne reakcie a prijať aj svoje slabé stránky. A v konfrontácii s realitou alebo s negatívnou skúsenosťou prichádza k psychickým problémom a kolapsu.[241] Ľudská bytosť je určitým spôsobom konfrontovaná

239 Napríklad Tim Kendall, bývalý vedúci pracovník Facebooku a následne prezident Pinterestu, hovorí o svojej vlastnej závislosti na technológiách Pinterestu, ktorých vývoj a zavádzanie sám viedol. Vedel o tom, presne poznal celú psychológiu a mechanizmy pôsobenia, no nevedel si v tom čase pomôcť. „Klasická irónia: počas dňa pracujem na niečom a tvorím niečo, čomu napokon sám podľahnem.“ „Je zaujímavé, že i napriek tomu, že som vedel, čo sa deje v zákulisí, nedokázal som svoje používanie sociálnych médií ovládať.“
ORLOWSKI, *The Social Dilemma*, 31:20, 31:55. [cit. 16. februára 2022].
Dostupné na internete: <https://www.netflix.com/sk/title/81254224>

240 ORLOWSKI, *The Social Dilemma*, 33:15 [cit. 16. februára 2022].
Dostupné na internete: <https://www.netflix.com/sk/title/81254224>

241 Podľa Dr. Jonathana Haidta, sociálneho psychológa z New York University, behom posledných desiatich rokov, ktoré sa prekrývajú s celosvetovým rozšírením využívania sociálnych sietí mladými ľuďmi a deťmi, prišlo k enormnému nárastu prípadov depresií, úzkosti a samovrážd medzi mladistvými a deťmi.
Napr. u dievčat (ktoré sú viac náchylné na ohodnotenie viditeľnej dokonalosti) nasledovne: u starších dievčat (15-19) stúplo seba poškodzovanie o 62%, u mladších (10-14) o 189%! Počet samovrážd

s neľudským perfekcionizmom algoritmov AI, ktorý je na hony vzdialený od empatie, emócií a nedokonalostí človeka, na ktoré však dokáže katastrofálne pôsobiť.[242]

Naviac, perimeter závislosti od pozitívneho hodnotenia nie je ohraničený reálnymi osobami z môjho okolia, ale rozširuje sa na tisíce až desiatky tisíc virtuálnych osôb, ktorých hodnotenie má na našu psychiku enormný dosah. Ľudský mozog sa nevyvinul tak, aby bol schopný prijímať spoločenské ohodnotenia v minútových intervaloch a od nesmierneho množstva iných subjektov. **Pod intenzívnym vplyvom sociálnych sietí a virtuálneho sveta sa ľudia stávajú závislí na svojej viditeľnej dokonalosti a neustálom prísune krátkodobých signálov odmeňovania až do tej miery, že si to spojujú s hodnotami a s pravdou.[243] Skutočne hodnotné a pravdivé sa tak stáva to, čo prináša najviac pozitívnych hodnotení a čo ľudí udržiava v kontrolovanom pozitívnom stave krátkodobých signálov odmeňovania. No nie je to zadanie ako vyšité pre systémy umelej inteligencie, ktoré bežia na pozadí najrozšírenejších sociálnych sietí?**

Vzhľadom na bezprecedentnú rozšírenosť týchto sociálnych sietí (a tým aj prakticky najrozšírenejšie využívanie systémov AI, ktoré interagujú s človekom a učia sa na reálnych dátach ľudí) ide o zásahy, ktoré ovplyvňujú celú modernú spoločnosť. Častokrát sa objavuje argument, že ide o krátkodobý výkyv, ktorý sprevádzal nástup akejkoľvek pokrokovej technológie v dejinách, resp. príchod ďalšej priemyselnej technológie a veľkej spoločenskej zmeny, a že je len otázkou času, kým si na to ľudstvo privykne. Dovolíme si tvrdiť – spolu s mnohými odborníkmi na etiku sociálnych sietí a informačnú spoločnosť –

narástol u starších dievčat o 70% a u mladších o 151%.
Ide o generáciu mladých, ktorí sa ako prví stretli so sociálnymi sieťami už na základnej škole, resp. skôr.
ORLOWSKI, *The Social Dilemma*, 40:00 [cit. 17. februára 2022].
Dostupné na internete: <https://www.netflix.com/sk/title/81254224>

242 Zaujímavé myšlienky a implikácie na tému dokonalosti AI uvádza článok Umelá inteligencia nesmie nahradiť nedokonalosť ľudskej empatie.
ADIB-MOGHADDAM, A. *Artificial intelligence must not be allowed to replace the imperfection of human empathy*. [on-line]. [cit. 20. februára 2022].
Dostupné na internete: <https://theconversation.com/artificial-intelligence-must-not-be-allowed-to-replace-the-imperfection-of-human-empathy-151636>

243 Por. vyjadrenia Chamatha Palihapatiyu, bývalého viceprezidenta Facebooku pre rast. Svojho času bol priekopníkom novátorských stratégií rastu, vďaka ktorým Facebook neskutočne rástol a ktoré sú v súčasnosti využívané naprieč technologickými firmami celého Silicon Valley.
ORLOWSKI, *The Social Dilemma*, 39:20 [cit. 17. februára 2022].
Dostupné na internete: <https://www.netflix.com/sk/title/81254224>

že v tomto prípade je to iné a to z dvoch dôvodov.

Realitou súčasnej digitálnej éry a ostatnej priemyselnej revolúcie je jej časová neohraničenosť – nachádzame sa v období permanentnej technologickej revolúcie, ktorá nie je uzavretým procesom, ktorého dôsledky spoločnosť vstrebáva. Ľudstvo je skôr vystavené neustálemu dopingu technológií, informácií a vplyvov systémov AI, ktorého ovocím je disproporcia medzi technikou a schopnosťou spoločnosti ju správne a bezpečne využívať.[244]

Druhým dôvodom je fakt, že informačné technológie stojace za týmito systémami sa nielen permanentne, ale predovšetkým exponenciálne zlepšujú. Na jednej strane interakcie je tak ľudský mozog, ktorý sa ďalej nevyvíja a na strane druhej technológie, ktoré sa za posledných šesťdesiat rokov mnohonásobne zmenili a porástli (len napríklad výkon počítačov sa za ten čas zvýšil v násobkoch miliárd).[245]

V konfrontácii uvedeného sa tak dotýkame niečoho, čo sme už skôr spomenuli ako technologickú singularitu[246] a čo je pre niektorých potvrdením obáv a výziev, ktoré obnáša

---

244 Por. ŠANTAVÝ, *Analýza virtuálneho sveta v kontexte náboženskej formácie, pôsobenia a života Cirkvi,* [on-line], s. 25, 29-30. [cit. 17. februára 2022].
Dostupné na internete: <https://peter.santavy.cloud/data/uploads/docs/analyza_virtualneho_sveta_v_kontexte_nabozenskej_formacie_posobenia_a_zivota_cirkvi.pdf>

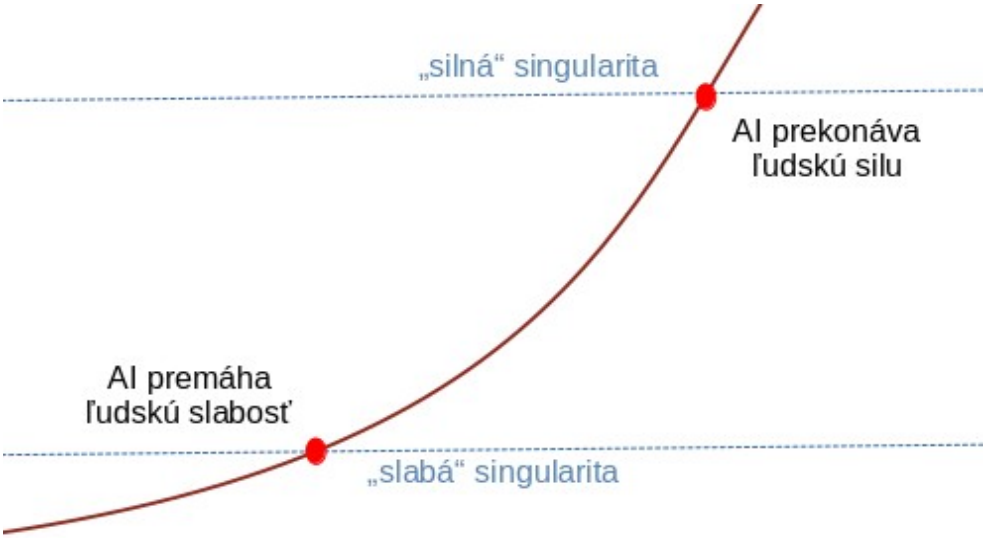
245 Por. Randima Fernando, spoluzakladateľ Centra pre humánne technológie, ORLOWSKI, *The Social Dilemma*, 45:00 [cit. 17. feburára 2022].
Dostupné na internete: <https://www.netflix.com/sk/title/81254224>
Por. ŠANTAVÝ, *Analýza virtuálneho sveta v kontexte náboženskej formácie, pôsobenia a života Cirkvi,* [on-line], s. 25, 29-30. [cit. 17. februára 2022].
Dostupné na internete: <https://peter.santavy.cloud/data/uploads/docs/analyza_virtualneho_sveta_v_kontexte_nabozenskej_formacie_posobenia_a_zivota_cirkvi.pdf>

246 Pod termínom technologická singularita sa myslí teoretický bod vo vývoji vedeckej civilizácie (znalostnej spoločnosti), v ktorom sa technologický pokrok zrýchli do nekonečna a prevýši všetky predpovede. *Singularita* [on-line]. [cit. 3. augusta 2020].
Dostupné na internete: <https://sk.wikipedia.org/wiki/Singularita>
Singularitou v umelej inteligencii je myslený stav, ktorý nastane, ak umelá inteligencia so schopnosťou učiť sa, zlepšovať sa a samostatne sa vyvíjať rýchlo dosiahne a následne prevýši inteligenciu človeka. Teda situácia, keď sa počítačové systémy stanú inteligentnejšími než ľudia.
Por. MITCHELL, *Artificial Intelligence*, s. 10.

koncept transhumanizmu[247], konkrétne vytvorením nových druhov ľudí genetickými manipuláciami alebo integráciu systémov AI a človeka pre eliminácie dôsledkov činnosti moderných systémov AI v službe sociálnych sietí a schopnosti spoločnosti ich asimilovať a integrovať.

Mnohí odborníci, pohybujúci sa v oblasti umelej inteligencie, podvedome čakajú na moment, keď umelá inteligencia premôže ľudskú silu a inteligenciu. Inak povedané, kedy nastane vek uvedomelej AGI, ktorá nás dokáže nahradiť v práci a bude múdrejšia než my. Stále sa domnievame, že táto otázka nie je na programe dňa a aktuálne je viac súčasťou pracovnej náplne futurológov.

Skôr sa – na základe uvedených rizík v tejto kapitole – stotožňujeme s pohľadom Tristana Harrisa, bývalého etika dizajnu vo firme Google a spoluzakladateľa Centra pre humánne technológie, pre ktorého je oveľa dôležitejší ten moment, v ktorom technológia prekoná a ovládne ľudské slabosti. Už vtedy prichádza víťazstvo AI a porážka ľudstva, lebo už vtedy prichádza závislosť, polarizácia a radikalizácia spoločnosti, zaslepenosť, strata schopnosti komunikovať a hľadať pravdu,... jednoducho prehráva všetko ľudské v nás...[248]



Obr. č. 10. Dva pohľady na singularitu v oblasti umelej inteligencie[249]

Míľnikom, ktorého by sme sa mali obávať, teda nie je budúca technologická singularita v oblasti umelej inteligencie, v ktorej AI prevýši náš intelekt, ale oveľa

247 O transhumanizme v kontexte umelej inteligencie sme pojednávali v kapitole 1.10.

248 ORLOWSKI, *The Social Dilemma*, 53:35 [cit. 17. februára 2022].
Dostupné na internete: <https://www.netflix.com/sk/title/81254224>

249 Dovolili sme si zaviesť obrazné termíny „slabá“ a „silná“ singularita (weak and strong singularity).

skôr moment, keď technológia ovládne a prekoná naše slabosti... už vtedy prichádza víťazstvo umelej inteligencie a porážka ľudstva.

V úvode tejto kapitoly sme spomenuli, že nutnou súčasťou extrémne rýchleho súčasného vývoja v oblasti umelej inteligencie je fungujúca možnosť monetizácie nasadenia systémov AI. **Pri riešení dôsledkov činnosti systémov AI je treba stále pamätať na to, aké ciele sú zadané pre činnosť týchto systémov**. Implementácia algoritmov s optimalizáciou na maximalizáciu zisku sa na sociálnych sieťach premieta do predlžovania času stráveného na sieti, počtu zhliadnutých reklám, množstva dát, ktoré používateľ vyprodukuje a ktoré je následne možné nejakým spôsobom premeniť na finančný benefit.

Samozrejme, takáto optimalizácia prináša aj svoje dôsledky: sociálne bubliny, v rámci ktorých sú podsúvané len tie informácie, ktoré korešpondujú s pohľadom používateľa a ponúkanie kontaktov na osoby rovnakého razenia, uprednostňovanie falošných informácií, pretože tie viac udržia používateľa pripojeného na sieti, ergo viac zarábajú.

Pre zadané ciele dnešné sociálne siete vytvárajú prostredie (t.j. algoritmy AI takto reagujú na zadané požiadavky a vylepšujú ponúkané informácie), ktoré je toxické - odtrhnuté od reality, s eróziou hodnôt, postmodernou[250] rezignáciou na hľadanie pravdy, zámenou dialógu za konfrontáciu,... Dôsledkom sociálnych sietí poháňaných sofistikovanými systémami AI je deštruktívne ovocie pre život a rozvoj človeka, kvalitu jeho života, vzťahy i celú spoločnosť. Jednoducho povedané – **zlyháva etický rozmer nasadenia AI**.

---

250 Treba si uvedomiť, že **informačný svet, zasadený do postmoderny s jej relativizáciou pravdy, rezignáciou na racio a akcentovaním emócií a zážitkov, je jedným z kľúčov k chápaniu výziev, rizík a potenciálu, ktorý v sebe vyššie uvedené problémy obnášajú**.
Virtuálny svet a sociálne siete sú tak dejiskom i katalyzátorom prerastania modernistickej racionalizácie etiky a morálnych hodnôt do postmodernistickej rezignácie na racio. Stráca sa schopnosť vnímať a hodnotiť veci i udalosti racionálne a objektívne. Čoraz väčší dôraz sa kladie na formu a zážitok, resp. emócie, ktoré sú prostredníctvom informačných technológií v mediálnom svete navodené. Človek následne častokrát hodnotí a rozhoduje sa pudovo alebo podvedome, povrchne, len na základe pozitívnej, či negatívnej emócie.
I keď badať reálny odstup od racionálneho vnímania, rozvíjanie postmoderny práve v rámci virtuálneho sveta a prostredníctvom prostriedkov IKT a systémov AI navodzuje falošnú atmosféru pseudovedeckých a pseudo hodnotových postojov, čo je živnou pôdou pre akékoľvek nekalé pôsobenie a spoločenskú, obchodnú i psychologickú manipuláciu s ľuďmi.
ŠANTAVÝ, P. *Niektoré výzvy informačnej spoločnosti v oblasti morálnej teológie*. In: *Doctorandum dies 2018 : Varia historia et moralia*. Bratislava: RKCMBF UK, 2018, s. 117-135.

Ak by sme sa na zmenu smerujúcu k etickému nasadeniu neodhodlali, vyhliadky môžu byť veľmi pochmúrne:[251]

- extrémne rozdelenie spoločnosti a narušenie vzťahov
- nárast zatvrdilej nevedomosti a ignorancie faktov
- neschopnosť riešiť aktuálne civilizačné výzvy
- premena demokracií na autokratické a disfunkčné celky
- zničenie globálnej ekonomiky a celosvetového spoločenstva
- potencionálne smerovanie k civilizačnému kolapsu...

Ako navrhované riešenie sa podsúva ponuka nových algoritmov umelej inteligencie a nástrojov, ktoré to vyriešia za nás – ako napr. uviedol Mark Zuckerberg, zakladateľ a CEO Facebooku (v súčasnosti Meta Platforms) pri vypočúvaní v Senáte v kauze ovplyvňovania prezidentských volieb v USA. V odbornej komunite však prevláda voči tomuto názoru skepsa – **vzhľadom na monetizačné zadanie, ktoré sa takmer vôbec v súčasnosti nedá zmeniť a schopnosti súčasných systémov AI (AI sama nevie rozoznať pravdu a mať etické mantinely) to musíme zmeniť my, ľudia.**

Ďalším dôvodom, prečo je táto zmena výzvou a povinnosťou pre človeka a spoločnosť, je skutočnosť, že **technológie samy o sebe nie sú tou priamou hrozbou - dokážu však prebudiť v človeku aj v spoločnosti tie najhoršie stránky ľudského bytia, ktoré sa tak stávajú existenčnou hrozbou**.[252]

Dôsledky sa teda dotýkajú celej spoločnosti a doliehajú aj na ľudí, ktorí sociálne siete nevyužívajú. Či už ide o znášanie „ovocia“ činov tých, ktorí sú týmito systémami priamo ovplyvnení, alebo sa stávajú obeťami spracúvania a využívania svojich osobných údajov zo sekundárnych informácií, ktoré o nich nazhromaždili automatizované systémy.

Znovu pripomíname, že pokiaľ bude existovať obchodný model, pre ktorý systémy AI v službe človeku a spoločnosti zlyhávajú, samo od seba sa to nezmení. V kontexte slabej umelej inteligencie (ANI) ešte stále konečnú voľbu cieľov nerobia stroje, ale človek, takže stále je v ľudskej moci tieto dôsledky ovplyvniť a zmeniť. Preto **musia existovať etické**

251 ORLOWSKI, *The Social Dilemma*, 1:20:25 [cit. 17. februára 2022]. Dostupné na internete: <https://www.netflix.com/sk/title/81254224>

252 ORLOWSKI, *The Social Dilemma*, 1:18:10 [cit. 17. februára 2022]. Dostupné na internete: <https://www.netflix.com/sk/title/81254224>

pravidlá a regulácie, ktoré jednotlivca i celú spoločnosť budú voči rizikám sofistikovanej práce systémov AI v rámci sociálnych sietí chrániť. Ak nechceme smerovať k dystopii, musíme sa preto na všetkých úrovniach angažovať a pracovať na tom spoločne ako ľudstvo.

V tomto duchu sa javia ako relevantné aj požiadavky:[253]

- **aby produkty a ciele, na ktoré sa systémy umelej inteligencie využívajú, boli humánne, t.j. pre dobro človeka**
- **aby tieto systémy a ich zadávatelia sa k nám nesprávali ako ku zdrojom, ktoré možno využívať a ťažiť z nich**

Viackrát sme sa v tejto kapitole okrajovo dotkli aj presahu k systémom uvedomelej AGI. K tomu ešte môžeme uviesť, že **vo vízii uvedomelej umelej inteligencie sa v odbornej komunite diskutuje aj ďalší špecifický rozmer závislosti – špecifické zameranie sa a naviazanie sa na pokročilú umelú inteligenciu ako na osobu.**

Okrem zhromažďovania a spracovania extrémneho množstva reálny dát o človeku, smerovania sociálnych sietí a závislostí spomeňme ešte niekoľko ďalších dôsledkov, na ktoré treba pamätať...

Jednou z ďalších zaujímavých oblastí aplikácie umelej inteligencie je vytváranie umelých virtuálnych identít (avatarov*)*, využívaných napríklad ako virtuálni hlásatelia v televíznych reláciách, alebo virtualizované identity reálnych hercov použité pri digitálnom nakrúcaní extrémnych scén moderných filmov. Schopnosti systémov AI v tejto oblasti sú obdivuhodné – veď takto dokážu „rozpohybovať“ nejednu súčasnú osobnosť, či vytvoriť umelé postavy na nerozoznanie od reálnych aktérov filmových dejov.

Nezávisle od tejto schopnosti tiež badať rozmach celej oblasti falošných (fake) obrázkov a videí zobrazujúcich reálne existujúce osoby v situáciách, na ktorých nemali účasť. Osobitne sa tento nešvár šíri v oblasti britkej internetovej zábavy, pornografie a v poslednom čase aj propagandy. Falošné fotografie a videá v tejto oblasti sú tým reálnejšie, čím viac zdrojov o danej osobe je k dispozícii. Napr. z jednej snímky bežného človeka zachytenej bezpečnostnou kamerou je veľmi náročné vytvoriť falošnú video scénu. Ale v prípade verejne známych osôb, napr. hercov alebo politikov, ktorých fotografií

253 ORLOWSKI, *The Social Dilemma*, 1:28:05 [cit. 17. februára 2022].
Dostupné na internete: <https://www.netflix.com/sk/title/81254224>

a videí sú plné profily sociálnych sietí, nie je problém, aby sa systém AI dostatočne natrénoval a vytvoril vynikajúce falzifikáty.[254]

Ak sa táto schopnosť spojí so skôr uvedenými systémami umelej inteligencie, ktoré dokážu vytvárať modely ľudského správania, tvoriť psychologické profily a predikovať správanie konkrétnych osôb, vo svete, ktorý sa stále viac orientuje na on-line komunikáciu a kybernetický priestor, prichádzajú na scénu mocné nástroje (**deepfake** systémy) na **vytváranie dôveryhodných virtuálnych identít ľudí a podvrhnutie ich falošnej komunikácie a konania** (samozrejme konania, ktoré sa nikdy reálne neuskutočnilo).[255]

Deepfake systémy sa podieľajú na stieraní hranice medzi pravdou a lžou, medzi realitou a fiktívnym svetom. **Veľmi reálne tak môže narastať nielen sociálna bublina, ktorej členovia uveria čomukoľvek (viď predchádzajúce state o rizikách sociálnych sietí), ale aj implicitná nedôvera k čomukoľvek ako základný postoj človeka pohybujúceho sa vo virtuálnom priestore.**

Ďalšou oblasťou dôsledkov nasadenia systémov umelej inteligencie sú právne aspekty súvisiace s ich využívaním.

Už doterajšia rozprava o závislostiach, zbere dát, monitorovaní, deepfake systémoch, atď. v sebe implicitne zahŕňala otázku legálnosti týchto systémov AI (napr. rozsah zhromažďovaných dát, podprahové ovplyvňovanie ľudí, atď.).

Z viacerých súčasných aktivít na tomto poli sa asi najďalej nachádza Európska únia so svojím regulačným rámcom zahŕňajúcim etické a legislatívne aspekty využívania systémov umelej inteligencie. Keďže ide asi o najdôležitejšiu reguláciu AI súčasnosti,

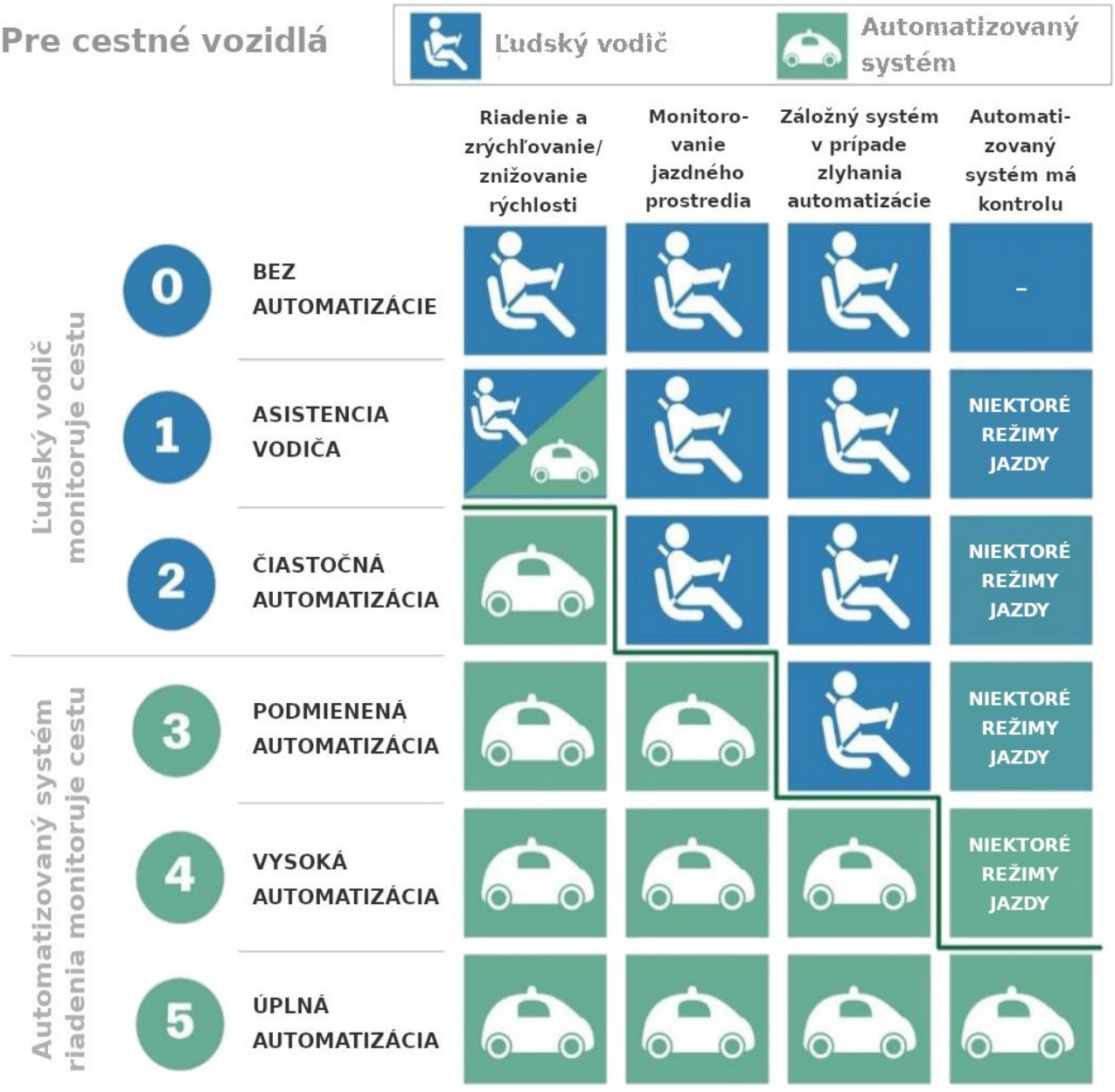
---

254 Ak si však myslíme, že pokiaľ nie sme verejne činnou osobou alebo celebritou, prípadne to nepreháňame na sociálnych sieťach, nám (aspoň v súčasnosti) tento problém nehrozí, k vytriezveniu by nás mohlo priviesť uvedomenie si dôsledkov pandemických obmedzení a hojného využívania práce z domu, ktorého súčasťou bývali aj nespočetné video stretnutia a porady na diaľku, napr. prostredníctvom platforiem Zoom alebo MS Teams. Mnohé z týchto videí boli ukladané a ich záznamy sa nachádzajú na dátových úložiskách spomenutých platforiem, resp. priamo konkrétnych firiem. Kompromitovanie týchto záznamov by okrem iného prinieslo stovky hodín kvalitného video materiálu, ktorý sa dá zneužiť na vytvorenie falošných videí.

255 Napr. obsah prejavu, ktorý by daný politik nikdy nepovedal; falošné videozábery udalostí, v ktorých vystupujú reálne osoby, ktoré však v skutočnosti nikdy aktérmi takej udalosti neboli; a pod. *It's Getting Harder to Spot a Deep Fake Video*. [on-line]. [cit. 21. marca 2022]. Dostupné na internete: <https://www.youtube.com/watch?v=gLoI9hAX9dw>

osobitne sa jej budeme venovať v tretej kapitole.

Klasickým príkladom právnych otázok spojených s využívaním umelej inteligencie v reálnom živote sú požiadavky kladené na autonómne vozidlá a zodpovednosť v prípade zapríčinenia dopravnej nehody.



Obr. č. 11. Päť úrovní automatizácie vozidla.[256]

Z pohľadu autonómie rozlišujeme päť stupňov – od vozidiel vybavených asistenčnými

256 AMUNDSEN, M. *The 5 levels of vehicle automation*. Twitter.
Dostupné na internete: <https://twitter.com/mamund/status/1171089135089700865>, upravené autorom.

službami až po plnú autonómiu riadenia, pričom až posledné dva stupne sú považované za tak samostatné, že pri riadení nevyžadujú priamu pozornosť tzv. bezpečnostného operátora (safety operator), t.j. človeka, ktorý dozoruje autonómny systém riadenia (viď obrázok č.11).

V súčasnosti sa diskutujú prvé návrhy na legislatívne požiadavky na bezpečnostného operátora, hľadajúc objektívnu mieru zodpovednosti ľudského faktora pri riadení autonómneho vozidla v rámci jednotlivých stupňov automatizácie.

Rieši sa dilema trénovania systémov AI plne autonómnych vozidiel pre osobitné kritické situácie, v ktorých sa musí autopilot rozhodnúť medzi potencionálnym ohrozením života posádky alebo ostatných účastníkov v čase blížiacej sa alebo prebiehajúcej dopravnej nehody.[257]

Vzhľadom na extrémnu technologickú komplexnosť autonómnych vozidiel vyvstáva aj otázka osobitných technologických noriem pre ich systémy AI a potrebnej cestnej (elektronickej) infraštruktúry (čo sme diskutovali v kapitole 2.4.) i akceptovanej miery neistoty, ktorá sprevádza autonómne systémy AI vzhľadom na limity a rizikové faktory, ktoré sme uviedli v kapitole 2.1.

V oficiálnych popisoch jednotlivých úrovní automatizácie vozidla sú autonómne funkcie vozidla pri úrovniach 1 až 4 podmienené vyjadrením „za určitých okolností/podmienok“.[258] Vozidlá piatej úrovne by mali by schopné plnej autonómnej jazdy za akýchkoľvek podmienok, čo však vylučuje nutnú potrebu cestnej a elektronickej infraštruktúry, osobitné podmienky prevádzky a pod. Na základe limitov a rizík, ktoré sme v kapitolách 2.1. až 2.4. uvádzali, však so súčasnými technológiami ANI nie je možné vozidlo piatej úrovne

257 Zaujímavým počinom v tejto oblasti je tzv. Moral Machine - on-line platforma vyvinutá na MIT, ktorá generuje morálne dilemy a zhromažďuje informácie o rozhodnutiach, ktoré ľudia robia medzi dvoma deštruktívnymi výsledkami.
Projekt Moral Machine, ktorý bol aktívny od januára 2016 do júla 2020, zhromažďoval ľudské postoje k morálnym riešeniam produkovaným systémami AI, napr. autonómnymi vozidlami. Prezentované scenáre a zozbierané informácie sú predmetom ďalšieho výskumu týkajúceho sa rozhodnutí, ktoré musí umelá inteligencia v budúcnosti robiť. Výskumné projekty ako Moral Machine pomáhajú nájsť riešenia pre náročné rozhodnutia o živote a smrti, ktorým budú čeliť autonómne vozidlá.
Moral Machine. [on-line]. [cit. 7. júla 2022].
Dostupné na internete: <https://www.moralmachine.net/>

258 Por. *Automated Vehicles for Safety*. [on-line]. [cit. 7. júla 2022].
Dostupné na internete: <https://www.nhtsa.gov/technology-innovation/automated-vehicles-safety>

realizovať. Totižto schopnosť samostatnej autonómnej jazdy vyžaduje použitie zdravého rozumu,[259] predovšetkým schopnosť pochopiť neznámu situáciu a navrhnúť riešenie na základe analógie s inými situáciami, ktoré sú pochopené v celkovom kontexte súvislostí, prípadne využitie tzv. základov intuície tak, ako ich uvádzame v kapitole 5.1.[260] Príchod skutočne autonómnych vozidiel piatej úrovne tak atakuje métu všeobecnej a silnej umelej inteligencie (AGI).

K diskusii dôsledkov využívania systémov umelej inteligencie pre spoločnosť treba doplniť oblasť, ktorá sa na prvý pohľad javí, že nie je na výslní, no je o to dôležitejšia – nasadenie algoritmov a systémov umelej inteligencie v oblasti spravodajských služieb, dohľadových systémov i algoritmického riadenia štátu a využitie v armáde.

2.6. Umelá inteligencia sa hlási do (spravodajskej) služby

Presnejšie povedané – umelá inteligencia v spravodajských službách už dávno nie je žiadny zelenáč a s technologickým rozvojom pracuje na svojom kariérnom postupe.

Umelá inteligencia nachádzala svoje uplatnenie v analytických sekciách spravodajských služieb už v osemdesiatych rokoch minulého storočia. V tom čase išlo primárne o využívanie symbolických systémov AI[261], reprezentovaných nástrojmi dolovania údajov (data mining[262]) a z dnešného pohľadu jednoduchého spracovania spravodajského obrazového materiálu. Primárne išlo o dáta získavané odposluchom z monitorovacích staníc rôzneho typu (od telefónnych operátorov, cez systémy rádio-elektronického boja až po Echelon[263]) a z fotografií získavaných prieskumným letectvom a družicami. Neskôr –

259 Por. MITCHELL, *Artificial Intelligence*, s. 268.

260 Ide o intuitívnu fyziku, biológiu a osobitne psychológiu, ktoré diskutujeme v kapitole 5.1.

261 Len pripomeňme, že symbolické systémy na základe definovaných pravidiel a postupov spracúvajú jednotlivé symboly (pojmy, slová, frázy,...) a vykonávajú priradené úlohy. Veľmi zjednodušene povedané – pomocou matematickej logiky sa snažia emulovať procesy myslenia.
Symbolické systémy sme popisovali v kapitole 1.4.

262 Data mining (dolovanie dát) je analytický proces navrhnutý na skúmanie veľkého množstva dát podľa konkrétnych vzorov a väzieb. Výsledkom sú množiny dát, spĺňujúcich žiadané podmienky, prípadne predikcia ďalšieho vývoja, správania, atď.
Data Mining Techniques [on-line]. [cit. 7. decembra 2015].
Dostupné na internete: <http://documents.software.dell.com/statistics/textbook/data-mining-techniques>

263 Ide o globálny systém pre zhromažďovanie a vyhodnocovanie odpočúvaných dát (SIGINT).
Z pôvodného sledovania vojenských a diplomatických cieľov počas studenej vojny sa ku koncu 20.

s rozvojom využívania internetu – sa pripojilo i monitorovanie globálnej počítačovej siete (išlo napr. o systém Carnivore[264]), ktoré postupne naberalo na dôležitosti úmerne tomu, ako sa v rámci rodiacej informačnej spoločnosti internet pretváral na virtuálny kybernetický priestor on-line života.

Preukázateľný profit z týchto systémov, ktorý presahoval vojenské a spravodajské záujmy a prinášal prvé ovocie aj v oblasti priemyselnej špionáže a globálneho dohľadu, dal vyrásť celému ekosystému spravodajských informačných nástrojov a ich pevnému zakoreneniu v spravodajských agentúrach. Súbežne s tým prichádzalo k prekračovaniu limitovaných technologických možností a lokálnych záujmov, čo dalo vznik globálnemu hromadnému dohľadu, t.j. hromadnému dohľadu nad celými populáciami a naprieč širokým svetom.[265]

Významným medzníkom pre celoplošné a detailné sledovanie boli dôsledky tragického teroristického útoku na Dvojičky v New Yorku z 11. septembra 2001. Podľa riaditeľa NSA, admirála Michaela S. Rogersa „útoky z 11. septembra podnietili 'zásadnú zmenu' v spôsobe, akým americká vláda využíva a zdieľa spravodajské informácie“.[266] **Bezpečnosť dostala absolútnu prioritu pred právom na súkromie**: USA Patriot Act prijatý v októbri 2001, účasť nadnárodných digitálnych korporácií na sledovacích

---

storočia rozvinul do celosvetového systému na zachytávanie i súkromnej i komerčnej komunikácie a stal sa tak nástrojom masívneho dohľadu a komplexnej špionáže.
Echelon [on-line]. [cit. 23. februára 2022].
Dostupné na internete: <https://en.wikipedia.org/wiki/ECHELON>

264 Carnivore bolo odpočúvacie a monitorovacie zariadenie FBI zamerané na emailovú a internetovú komunikáciu. Aktívne využívané bolo v rokoch 1997 až 2005, následne bolo nahradené výkonnejšími a globálnymi dohľadovými systémami.
Carnivore [on-line]. [cit. 23. februára 2022].
Dostupné na internete: <https://en.wikipedia.org/wiki/Carnivore_(software)>

265 Okrem Echelonu a Carnivore môžeme hovoriť o takých systémoch, ako sú XKeyscore, PRISM, Dishfire, Stone Ghost, Tempora, Frenchelon, Fairview, MYSTIC, DCSN, Boundless Informant, Bullrun, Pinwale, Stingray, SORM, RAMPART-A, Mastering the Internet a Jindalee Operational Radar Network.
Global surveillance [on-line]. [cit. 23. februára 2022].
Dostupné na internete: <https://en.wikipedia.org/wiki/Global_surveillance>

266 GARAMONE, J. *9/11 Drove Change in Intelligence Community, NSA Chief Says*. [on-line]. [cit. 27. februára 2022].
Dostupné na internete: <https://www.defense.gov/News/News-Stories/Article/Article/945544/911-drove-change-in-intelligence-community-nsa-chief-says/>

programoch,[267] kontroly počítačovej techniky rôznych oficiálnych predstaviteľov štátu pracovníkmi CIA (2014), odhalenie sofistikovaných nástrojov používaných na prienik do smartfónov, počítačov a dokonca aj do „smart“ televíznych prijímačov pripojených na internet,[268] integrácia systémov umelej inteligencie na rozpoznávanie tvárí (Clearview AI) v roku 2020, atď.[269]

Zmena priority zo súkromia na národnú bezpečnosť v USA spustila celý sled udalostí, ktoré sa ako domino efekt šírili aj v ostatných krajinách: implementácia moderných sledovacích systémov štátnymi orgánmi, ich využívanie na lokálnej úrovni samospráv i v komerčnej sfére a – čo je najhoršie – bol umožnený neriadený prístup k týmto technológiám aj silovým a spravodajským zložkám autoritárskych režimov a diktatúr.

Súčasné využitie systémov umelej inteligencie v dohľadových systémoch a pri sledovaní je ovplyvnené dvomi skutočnosťami:

- spoločenským tlakom a zmenou politickej klímy po kauze Snowden vs. NSA
- veľkým pokrokom v rozvoji systémov AI ako systémov hlbokého strojového učenia

Edward Snowden je americký whistle-blower[270], ktorý v roku 2013 dal novinárom denníka The Guardian k dispozícii dokumenty o sledovacích aktivitách americkej National Security Agency (NSA) či britskej Government Communications Headquarters (GCHQ). Dovtedy pracoval pre NSA na rôznych pozíciách ako špecialista na informačné technológie a spoznajúc praktiky s nelegálnymi dohľadovými systémami sa rozhodol tieto informácie

267 Microsoft sa údajne v roku 2007 stal prvou veľkou technologickou spoločnosťou, ktorá spolupracovala na programe elektronického sledovania občanov PRISM. Neskôr sa mali pripojiť i Yahoo, Google a Facebook.

268 *WikiLeaks Releases Trove of Alleged C.I.A. Hacking Documents*. [on-line]. [cit. 27. februára 2022]. Dostupné na internete: <https://www.nytimes.com/2017/03/07/world/europe/wikileaks-cia-hacking.html>

269 *How 9/11 Changed the Course of Personal Data Collection and Surveillance*. [on-line]. [cit. 27. februára 2022].
Dostupné na internete: <https://www.startpage.com/privacy-please/startpage-articles/how-9-11-changed-the-course-of-personal-data-collection-and-surveillance>

270 Výraz whistleblowing sa označuje zverejnenie nekalých, resp. nelegitímnych praktík spoločností alebo štátnych inštitúcií interným zamestnancom. Ide buď o zverejnenie mediálne, alebo o upozornenie konkrétnych štátnych orgánov.
Por. *Whistleblowing* [on-line]. [cit. 23. februára 2022].
Dostupné na internete: <https://cs.wikipedia.org/wiki/Whistleblowing>

z morálnych dôvodov zverejniť.[271]

Využívajúc šifrovanie, prostriedky darknetu[272] a profesionálne krytie, Snowden sa tajne stretol s redaktormi The Guardian, ktorí 5. júna 2013 zverejnili prvý z článkov založených na jeho odhalení.[273]

V priebehu nasledovných mesiacov boli z množstva tajných informácií, ktorými Snowden disponoval, zverejnené nasledovné kauzy:[274]

- tajný súdny príkaz, ktorým americká vláda donútila telekomunikačnú firmu Verizon Communications sprístupniť milióny záznamov telefonických hovorov[275]
- tajný národný bezpečnostný program elektronického sledovania občanov PRISM, ktorý NSA prevádzkovala v USA od roku 2007[276]
- odhalenie tajného programu Tempora z dielne britskej spravodajskej služby GCHQ, ktorý od roku 2012 neustále sleduje a aj fyzicky ukladá telefónne hovory (max. 30 dní), internetové informácie a metadáta ľubovoľných osôb na svete, t.j. priamo ich

---

271 „Uvedomil som si, že som súčasťou mašinérie, ktorá spôsobuje viac zlého než dobrého.“ *Profile: Edward Snowden.* In: *BBC*. [on-line]. 2013, 24. 6. [cit. 23. februára 2022]. Dostupné na internete: <http://www.bbc.co.uk/news/world-us-canada-22837100>

272 Darknet je technologicky izolovaná časť internetu a IKT vo všeobecnosti, na ktorej je kladený dôraz na anonymitu a bezpečnosť. Jej možnosti sú najčastejšie využívané buď na ochranu pred sledovaním, alebo na kriminálne aktivity. Darknet je častokrát spojovaný s jeho obsahom (Darkweb) a skrytými službami (Tor services).
Por. *Darknet* [on-line]. [cit. 23. februára 2022].
Dostupné na internete: <https://en.wikipedia.org/wiki/Darknet>

273 GIDDA, M. *Edward Snowden and the NSA files – timeline* [on-line]. The Gurdian. [cit. 23. februára 2022].
Dostupné na internete: <http://www.guardian.co.uk/world/2013/jun/23/edward-snowden-nsa-files-timeline>
ŠANTAVÝ, *Niektoré výzvy informačnej spoločnosti v oblasti morálnej teológie, s*. 117-135.

274 ŠANTAVÝ, *Niektoré výzvy informačnej spoločnosti v oblasti morálnej teológie, s*. 117-135.

275 Vzhľadnom na ústavu USA a jej dodatky išlo o nezákonné a nelegitímne konanie voči občanom USA. GIDDA, *Edward Snowden and the NSA files – timeline* [on-line]. [cit. 23. februára 2022].
Dostupné na internete: <http://www.guardian.co.uk/world/2013/jun/23/edward-snowden-nsa-files-timeline>

276 Por. *PRISM (NSA)* [on-line]. [cit. 23. februára 2022].
Dostupné na internete: <https://sk.wikipedia.org/wiki/PRISM_(NSA)>

telefónne hovory, obsahy ich e-mailových správ, ich príspevky na Facebooku, zoznam nimi navštívených stránok atď. GCHG získané dáta a analýzy poskytuje aj NSA.[277]

- odhalenie odpočúvania čínskych hovorov a sms agentúrou NSA[278]
- odhalenie odpočúvania a napojenia sa na počítačové siete, ktoré NSA konalo v budovách diplomatického zastúpenia EÚ a OSN vo Washingtone a pri OSN. V dokumentoch z roku 2010 sú Európania explicitne uvedení ako cieľ sledovacieho útoku.[279]

Snowdenove odhalenia boli sprevádzané veľkou mediálnou odozvou. To, pred čím niektorí odborníci z oblasti informačnej bezpečnosti pre rokom 2013 varovali a čo väčšina odbornej verejnosti i politikov považovala za prehnané či paranoidné, sa ukázalo ako pravdivé: **prostriedky a možnosti informačných technológií a systémov umelej inteligencie v spoločnosti, ktorá sa mení na informačnú, sú reálne zneužívané tajnými službami, vládnymi agentúrami a korporáciami**.[280]

Okrem morálneho a etického hodnotenia zverejnenia prísne tajných dokumentov spravodajských služieb USA sa doba po Snowdenovi snaží pragmaticky vyrovnať s odhalenými problémami a v súčasnosti sa už – technologicky, spoločensky i politicky – úplne inak pristupuje k ochrane súkromia a bezpečnosti dát. Jednak rastie povedomie o práve na súkromie, no súbežne sa vo viacerých krajinách stupňuje tlak na legalizáciu

277 Por. *Tempora* [on-line]. [cit. 23. februára 2022].
Dostupné na internete: <https://sk.wikipedia.org/wiki/Tempora>

278 LAM, L., CHEN, S. *US spies on Chinese mobile phone companies, steals SMS data: Edward Snowden*
In: *South China Morning Post.* [on-line]. 2013, 23. 3. [cit. 23. februára 2022].
Dostupné na internete: <https://www.scmp.com/news/china/article/1266821/us-hacks-chinese-mobile-phone-companies-steals-sms-data-edward-snowden>

279 POITRAS, L., ROSENBACH, M., SCHMID, F., STARK, H. *NSA horcht EU-Vertretungen mit Wanzen aus*
In: *Der Spiegel.* [on-line]. 2013, 29. 6. [cit. 23. februára 2022].
Dostupné na internete: <http://www.spiegel.de/netzwelt/netzpolitik/nsa-hat-wanzen-in-eu-gebaeuden-installiert-a-908515.html>
POITRAS, L., ROSENBACH, M., STARK, H. *NSA überwacht 500 Millionen Verbindungen in Deutschland*
In: *Der Spiegel.* [on-line]. 2013, 30. 6. [cit. 23. februára 2022].
Dostupné na internete: <http://www.spiegel.de/netzwelt/netzpolitik/nsa-ueberwacht-500-millionen-verbindungen-in-deutschland-a-908517.html>

280 ŠANTAVÝ, *Niektoré výzvy informačnej spoločnosti v oblasti morálnej teológie*, s. 117-135.

odhalených metód špehovania a monitorovania ľudí, aby konanie, ktoré bolo predtým tajné a v súčasnosti je mnohokrát technologicky blokované[281], bolo legalizované a nariadené zákonom.[282] Základným argumentom pre legalizáciu dohľadu a špehovania býva potreba národnej bezpečnosti (kybernetický terorizmus a terorizmus vo všeobecnosti, vojenská obrana štátu a boj s dezinformáciami) a boj s kriminalitou (detská pornografia, obchodovanie s ľuďmi a drogová činnosť).[283]

S rastom povedomia o ochrane súkromia i bezpečnostných hrozbách v spoločnosti a zlepšujúcou sa ochranou informačných prostriedkov i širokou dostupnosťou bezpečnostných nástrojov a návodov by sa mohlo zdať nebezpečenstvo zneužívania technológií na sledovanie obyvateľov zažehnané (ak si teda odmyslíme v poslednom období rastúci tlak na legalizáciu).[284]

Opak je však pravdou – na scénu prichádza veľký pokrok v rozvoji systémov umelej inteligencie. Súčasné algoritmy hlbokého učenia totižto vedia skĺbiť dôsledky spoločenského tlaku, v mnohých prípadoch pretaveného do legislatívnych obmedzení, s dostupnosťou extrémneho množstva dát, ktoré – ako inak – sú vytvárané ľudmi, teda potencionálnymi subjektami sledovania. Tieto systémy využívajú dve skutočnosti:

- vďaka svojej schopnosti autonómne fungovať a adaptabilite[285] tieto systémy dokážu spracúvať dáta, ku ktorým bez príkazu súdov neakceptujeme prístup osôb. To, že výsledná analýza a predložené závery sú prijímané s vysokou mierou dôveryhodnosti, je už len čerešnička na torte.
- spracúvanie metadát, ktoré netvoria priamo obsah komunikácie. Metadáta sú informácie o komunikácii – kto, kedy, komu, ako častokrát, v akej korelácii s inými

281 Zlepšujúca sa ochrana systémov voči sledovaniu a široká dostupnosť bezpečnostných nástrojov.

282 Svojho času vyvolal veľký rozruch návrh zákona v Austrálii, podľa ktorého môže byť vlastník zašifrovaného zariadenia odsúdený až na desať rokov väzenia, ak odmietne odšifrovať svoje zariadenie. Už v tom čase podobnú legislatívu mala napr. Veľká Británia.
Vo viacerých krajinách je zasa povinnosťou cestujúceho pri vstupe do krajiny odblokovať svoje zariadenie a umožniť jeho preverenie úradom.

283 ŠANTAVÝ, *Niektoré výzvy informačnej spoločnosti v oblasti morálnej teológie,* s. 117-135.

284 Radi by sme upozornili čitateľa na postupné rozširovanie záberu tejto kapitoly – od spravodajských služieb k sledovaniu a dohľadu vo všeobecnosti...

285 Vďaka patrí nielen moderným algoritmom a špecialistom na hyper parametre, no predovšetkým nám všetkým za poskytnuté množstvo údajov a dát:-)

udalosťami a osobami, atď.[286]

V roku 2014 bývalý riaditeľ CIA a NSA na The Johns Hopkins Foreign Affairs Symposium vyhlásil: „na základe metadát zabíjame ľudí“.[287]

Systémy umelej inteligencie, ktorým sa v súčasnosti analýza metadát zveruje, tak de facto majú rastúci potenciál rozhodovať o bytí a nebytí, „generovať“ dôkazy, ktoré môžu byť použité napríklad pri obvinení z vlastizrady, identifikácii podozrivých osôb, dokazovaní nelegálnej činnosti a pod.

Ak si uvedené spojíme s rizikami a limitmi komunikovanými v kapitolách 2.1. - 2.4. a okoreníme to niektorými spoločenskými dôsledkami z kapitoly 2.5., nezostáva nám nič iné, len **akcentovať veľkú opatrnosť a vyžadovať striktnú zákonnosť i dôsledný dohľad demokraticky zvolených zástupcov pri tomto druhu nasadenia systémov AI**.

V úvode tejto kapitoly sme trošku nadnesene spomenuli „kariérny postup“ umelej inteligencie v spravodajstve... Kombinácia extrémne veľkých objemov dát (big data) s modernými algoritmami umelej inteligencie v kontexte informačnej spoločnosti i napriek mnohým legislatívnym a regulačným obmedzeniam zabezpečuje systémom AI dôležité postavenie medzi prostriedkami spravodajských služieb a svojmu kariérnemu postupu dáva „zelenú“.

Ak by sme sa ponorili do sveta spracovania dát i metadát systémami AI, do sveta technológií priebežne rozpoznávajúcich tváre a identifikujúcich osoby, žasli by sme nad aplikačnými možnosťami, ktoré tento svet ponúka a možno by sme zistili, aké pokušenie to okrem legálnych a prospešných scenárov využitia prináša... A odtiaľ už je len krok k ešte väčšiemu rozšíreniu týchto systémov v spoločnosti, než sme uvádzali ako dôsledok zmien bezpečnostných priorít po páde Dvojičiek – či už v pozitívnom alebo v negatívnom slova zmysle.

V súčasnosti badať neustále rastúci počet štátov a korporácií, ktoré zavádzajú pokročilé

---

286 Napr. pri emailovej komunikácii obsah emailu tvorí dáta. Metadátami sú napr. emailová adresa prijímateľa, čas odoslania emailu, ďalší prijímatelia na kópii, interval medzi prijatím emailu a odpoveďou, korelácia medzi emailovými komunikáciami viacerých sledovaných osôb, GPS súradnice miesta, z ktorého bol email odoslaný,....

287 „We kill people based on metadata.“ *The Price of Privacy: Re-Evaluating the NSA* [on-line]. [cit. 23. februára 2022]. Dostupné na internete: <https://www.youtube.com/watch?v=kV2HDM86XgI&t=18m>

nástroje umelej inteligencie umožňujúce monitorovanie, dohľadávanie a sledovanie občanov. Jednou z výpovedných štúdií popisujúcej vývoj v tejto oblasti je *The Global Expansion of AI Surveillance* a *AI Global Surveillance (AIGS) Index* z dielne nadácie Carnegie Endowment for International Peace.[288]

Autor štúdie, v snahe uchopiť a mať možnosť riešiť dôsledky nasadenia pokročilých nástrojov sledovania využívajúcich umelú inteligenciu, sa zameriava na scenáre ich použitia, t.j. ako sa tieto nástroje nasadzujú a ako sa používajú. Výsledkom je Index globálneho dohľadu s umelou inteligenciou (AIGS) zhromažďujúci empirické údaje o používaní dohľadu systémami AI v 176 krajinách sveta. Index i celá štúdia v zásade nerozlišujú medzi legitímnym a nelegitímnym používaním dohľadu systémami AI – cieľom je skôr ukázať, ako nové možnosti dohľadu menia schopnosť vlád monitorovať a sledovať jednotlivcov alebo systémy. Štúdia uvádza krajiny, ktoré využívajú technológiu dohľadu pomocou umelej inteligencie, sumarizuje konkrétne typy dohľadu, ktoré sú vládami zavádzané a vymenúva krajiny i spoločnosti dodávajúce túto technológiu.

Ku kľúčovým zisteniam štúdie[289] patrí prekvapivá rýchlosť šírenia implementácie dohľadových systémov AI v rámci širokého spektra štátov. Zo 176 vymenovaných krajín minimálne sedemdesiatpäť využíva tieto prostriedky naozaj aktívne a intenzívne.

K najviac používaným technológiám umelej inteligencie v dohľadových systémoch patria platformy tzv. inteligentných, resp. bezpečných miest (päťdesiatšesť krajín), systémy rozpoznávania tváre (šesťdesiatštyri krajín) a inteligentná polícia (päťdesiatdva krajín).

Hlavným aktérom na poli vývoja, ale i zavádzania týchto technológií do praxe je Čína. Jej ponuka je naviac veľmi atraktívna – s veľkými zľavami, či pomocou zvýhodnených pôžičiek umožňuje zavádzanie svojich systémoch i v krajinách a režimoch, ktoré by si inak podobné systémy vôbec nemohli dovoliť.

Ďalšími významnými hráčmi na trhu dohľadových systémov umelej inteligencie sú Japonsko, USA, Francúzsko, Nemecko a Izrael. **Je znepokojujúce, že demokratické**

288 FELDSTEIN, S. *The Global Expansion of AI Surveillance*. [on-line]. [cit. 28. februára 2022]. Dostupné na internete: <https://carnegieendowment.org/2019/09/17/global-expansion-of-ai-surveillance-pub-79847>

289 FELDSTEIN, *The Global Expansion of AI Surveillance*. [on-line]. [cit. 28. februára 2022]. Dostupné na internete: <https://carnegieendowment.org/2019/09/17/global-expansion-of-ai-surveillance-pub-79847>

štáty, z ktorých väčšina týchto systémov pochádza, neprijímajú primerané opatrenia na monitorovanie a kontrolu šírenia sofistikovaných technológií spojených s celým radom možných porušení ľudských práv a zneužití autokratickými režimami a diktatúrami.

Demokratické štáty sú hlavnými používateľmi dohľadu pomocou umelej inteligencie. Z AIGS vyplýva, že 51% vyspelých demokracií využíva systémy dohľadu s umelou inteligenciou. Ide o celý rad technológií dohľadu, od platforiem bezpečného mesta až po kamery na rozpoznávanie tváre. Javí sa, že miera zneužitia je nízka, pretože najdôležitejším faktorom, ktorý určuje, či vlády nasadia túto technológiu na represívne účely, je kvalita a transparentnosť ich vládnutia spojené s regulačnými a legislatívnymi štandardmi.

Vlády v autokratických a semi-autokratických krajinách sú náchylnejšie na zneužívanie dohľadových technológií ako vlády v demokraciách. Využívajú systémy AI na účely masového dohľadu, na posilnenie represie i na dosiahnutie určitých politických cieľov.

Existuje silný vzťah medzi vojenskými výdavkami krajiny a tým, ako vláda využíva dohľadové systémy AI: štyridsať z päťdesiatich krajín s najvyššími vojenskými výdavkami na svete (na základe kumulatívnych vojenských výdavkov) využíva aj technológiu dohľadu pomocou AI.

Toľko z citovanej správy a indexu globálneho dohľadu.

Môžeme len doplniť, že kým v niektorých krajinách rastie snaha o transparentné a regulované využívanie systémov AI aj v oblasti dohľadu[290], v iných prináša nasadenie algoritmov umelej inteligencie vyšší stupeň monitorovania a dohľadu nad jednotlivcami i nad celou spoločnosťou. Znepokojivým príkladom tohoto trendu je Čína so svojím systémom sociálneho kreditu[291], ktorý pomocou sofistikovaného dohľadového systému, analýzy širokého spektra dát a osobných údajov hodnotí a reguluje činnosť osôb a spoločností v štáte. Tento systém intenzívne využíva technológie umelej inteligencie s cieľom získať efektívnejšie nástroje riadenia správania. Ovocím je bezprecedentný dohľad nad akýmikoľvek aktivitami občanov, ich kategorizácia a obmedzovanie, ak nespĺňajú štátom požadované „parametre“. Ide o nebezpečný

290 Napríklad aktuálne uvedený regulačný rámec EÚ pre systémy umelej inteligencie.

291 Schematické znázornenie systému sociálneho kreditu je uvedené v prílohe.

precedens, ktorý sa v Číne naďalej rozvíja a – žiaľ – stáva sa inšpiratívnym pre predstaviteľov viacerých krajín (aj demokratického) sveta.[292]

Pozornému čitateľovi iste neuniklo, že sme – vzhľadom na kontroverznosť a dôsledky sledovania – komunikovali nasadenie algoritmov umelej inteligencie prevažne v rámci dohľadových systémov. No niektoré závery, ktoré sme uviedli zo štúdie The Global Expansion of AI Surveillance nám pripomínajú aj ďalšie dva dôležité rozmery spravodajského nasadenia – analytiku a predikcie.

Plošný zber dát, sledovacie a dohľadové systémy by bez analytickej nadstavby boli nefunkčné. Nie je totiž v ľudských silách spracúvať extrémne množstvo spravodajských dát a už vôbec nie sme schopní nad týmito dátami vytvárať pokročilé analýzy a predikcie.

Práve preto bola jednou z prvých vo využití ešte symbolických systémov AI oblasť Data Miningu (dolovanie dát), ktorá – ako sme uviedli v úvode tejto kapitoly – predstavuje analytické metódy a procesy navrhnuté na skúmanie veľkého množstva dát podľa konkrétnych vzorov a väzieb. Výsledkom sú množiny dát, splňujúcich žiadané podmienky, prípadne predikcia ďalšieho vývoja, správania, atď.

S rozvojom algoritmov hlbokého učenia, fenoménu big data, metadát a kybernetického priestoru data miningové systémy dostali nový dych – sú oveľa sofistikovanejšie, výsledky ponúkajú v reálnom čase a zo vstupných dát vedia vydolovať oveľa viac.

Zároveň – a tu by sme odkázali na postrehy z kapitoly 2.5. ohľadom schopnosti systémov AI vytvárať modely, ktoré sú schopné predpovedať naše konanie – kvalitná analýza je nutnou podmienkou pre modelovanie faktorov, na ktoré sa dohľad a analýza zameriavajú.[293]

Trio technológií umelej inteligencie – sledovanie a dohľadové systémy spojené s analytickými nástrojmi a schopnosťou vytvárať modely predikujúce ľudské konanie či vývoj situácie – tvorí v súčasnosti skutočne silnú technologickú výbavu (nielen) spravodajských služieb.

V kontexte tejto technologickej výbavy a čínskeho systému sociálneho kreditu vráťme sa ešte k zisteniam štúdie *The Global Expansion of AI Surveillance*. Konkrétne k najviac

292 *Social Credit System*. [on-line]. [cit. 28. februára 2022]. Dostupné na internete: <https://en.wikipedia.org/wiki/Social_Credit_System>

293 Napríklad modelovanie veľkosti rizika a parametrov teroristického útoku v danej oblasti, vývoj kriminality v konkrétnej časti veľkomesta,...

používaným technológiám AI v dohľadových systémoch, medzi ktoré patria platformy tzv. inteligentných, resp. bezpečných miest, systémy rozpoznávania tváre a inteligentná polícia. Tieto platformy sú súčasťou technológií, známych pod názvami ako **algokracia**, vláda podľa algoritmov, algoritmický právny poriadok, algoritmické vládnutie a pod.

Ide o alternatívnu formu vlády, resp. spoločenského usporiadania, pri ktorom sa na reguláciu, presadzovanie práva a všeobecne na akýkoľvek aspekt každodenného života, ako je doprava alebo registrácia pozemkov, používajú počítačové algoritmy, najmä umelá inteligencia a blockchain.

Algoritmické vládnutie na jednej strane vďaka použitým technológiám môže priniesť nezanedbateľné benefity v správe vecí verejných, skvalitnení a zefektívnení služieb občanom i v riadení štátu, na druhej strane však prináša rôznorodé výzvy a riziká. K výzvam sa určite radí potreba harmonizácie algoritmického vládnutia s princípmi riadenia moderného štátu a e-governmentu i zvládnutie sociologických aspektov a praktických dôsledkov algokracie.

Ako sme videli na záveroch správy nadácie Carnegie Endowment for International Peace, nasadenie systémov AI smerujúce k algokracii sprevádza viacero problémov v oblasti demokracie a ľudských práv. Nesmieme však zabúdať na skutočnosť, že medzi riziká algoritmického vládnutia patrí aj celý komplex problémov prameniacich z limitov a rizikových faktorov súčasných systémov AI, ktoré sme uvádzali v kapitolách 2.1. až 2.5.

Sumarizujúc túto kapitolu vnímame, že technológie využívajúce algoritmy umelej inteligencie dokážu byť veľmi účinným nástrojom pre spravodajské služby a dohľad s takým presahom do ostatných oblastí verejného života a štátu, ktorý môže znamenať veľký posun v procesoch a spôsobe fungovania spoločnosti. Avšak vzhľadom na politický, ideologický i spoločenský kontext v rôznych častiach sveta a riziká systémov AI treba zabezpečiť, aby prostriedky umelej inteligencie neboli zneužité alebo sa vymkli kontrole.

Preto je treba v celom spektre nasadenia od spravodajských služieb až po algoritmické riadenie akcentovať veľkú rozvážnosť a vyžadovať striktnú zákonnosť i dôsledný dohľad demokraticky zvolených zástupcov, obmedziť dopady na sociálnu spravodlivosť[294] a zabezpečiť dodržiavanie ľudských práv a hodnôt.

294 Digital divide – digitálne rozdelenie ako dôsledok informačnej nerovnosti, ktorá prerastá do digitálnej chudoby a má evidentné dôsledky na život ľudí. Digital divide spomíname v 3. i 4. kapitole.

2.7. Systémy umelej inteligencie narukovali do armády

Na prelome osemdesiatych a deväťdesiatych rokov minulého storočia sme ešte pred nežnou revolúciou ako študenti technickej kybernetiky žartovali, kedy bude u nás prvý 32-bitový procesor. Odpoveď: s príletom prvej riadenej strely zo západu.[295]

Do akej miery sa situácia zmenila, vyjadruje skutočnosť, že v súčasnosti už vôbec neriešime, akú procesorovú jednotku moderný zbraňový systém obsahuje, ale hodnotíme, akou technológiou umelej inteligencie disponuje. A nielen „na západe“...

V kapitole 1.8. sme uviedli, že v tzv. zimných časoch umelej inteligencie, t.j. v obdobiach veľkého útlmu na poli AI bol vývoj udržiavaný v zásade len v rámci základného výskumu viacerých univerzitných centier a pokroku v príbuzných oblastiach (napr. robotika a kybernetika, výpočtová technika, dátová a počítačová veda, v rámci priemyslu a vojenského vývoja[296]).

Vzhľadom na predikované možnosti umelej inteligencie už od čias Dartmouthského seminára stál rozvoj vojenskej techniky a armádne nasadenie ako jeden z prvých čakateľov v rade na reálne využitie. A akonáhle ostatná jar AI prerástla do prakticky kontinuálneho rozvoja v tejto oblasti,[297] umelá inteligencia sa stala jednou z primárnych technológií nielen moderných zbraňových systémov, ale vojenského využitia v celej svojej šírke a komplexnosti.

Využívanie technológií AI v armáde nie je zviazané len s krajinami, u ktorých by sme to vzhľadom na vedecko-technologický, či vojenský potenciál predpokladali. S dostupnosťou moderných systémov založených na umelej inteligencii a ich komercializáciou sú tieto systémy – ako sme uviedli pri dohľadových systémoch v predchádzajúcej kapitole –

Por. ŠANTAVÝ, *Analýza virtuálneho sveta v kontexte náboženskej formácie, pôsobenia a života Cirkvi,* [on-line], s. 30-32. [cit. 28. februára 2022].
Dostupné na internete:
<https://peter.santavy.cloud/data/uploads/docs/analyza_virtualneho_sveta_v_kontexte_nabozenskej_formacie_posobenia_a_zivota_cirkvi.pdf>

295 Išlo o riadené strely BGM-109 Tomahawk a MGM-31B Pershing II.

296 Veľkú roľu v tejto oblasti dlhodobo zohráva **DARPA** (Defense Advanced Research Projects Agency) – agentúra ministerstva obrany USA zodpovedná za výskum a vývoj nových vojenských technológií. Dostupné na internete: <https://www.darpa.mil/>

297 Eventuálny postupný prerod striedania ročných období AI do prakticky kontinuálneho rastu a rozvoja sme diskutovali v kapitole 1.8.

lákadlom prakticky pre kohokoľvek.[298]

Využívanie systémov AI vo vojenskej oblasti napreduje predovšetkým v týchto krajinách:

- veľmoci (USA, Čína, Rusko,...), ktoré majú i dostatočný vedecko-technologický potenciál i rozsiahle armádne celky
- vysoko technologicky rozvinuté štáty (Izrael, Japonsko, niektoré štáty EÚ,...), ktorých technologické portfólio takmer prirodzene zahŕňa aj armádne využitie
- vysoko militarizované krajiny (India, Turecko,...), ktoré budujú moderné armádne celky a investujú do najmodernejších technológií v tejto oblasti
- problematické štáty (Irán, Severná Kórea,...), ktoré pre svoje ideologické a politické ciele kladú dôraz na rozvoj vojenského potenciálu, no v rámci svojich možností sa zameriavajú na pre nich dostupné a zároveň efektívne technológie, ku ktorým patria aj systémy AI

Uvedené krajiny sa snažia uchopiť využívanie umelej inteligencie vo vojenskej oblasti komplexne. V súčasnosti však nielen ony, ale takmer každá krajina, ktorá sa snaží rozvíjať, resp. budovať moderné armádne zložky, využíva systémy AI aspoň v niektorej z oblastí, medzi ktoré patrí:

- vojenské spravodajstvo
- modelovanie technológií, konfliktov a operácií
- podpora pre velenie
- trenažéry, simulátory a výcvik
- autonómne zbraňové systémy
- skupinové riadenie bojových prostriedkov a autonómnych systémov

298 Už v roku 2019 vtedajší minister obrany USA Mark Esper predložil informácie o čínskych exportných aktivitách, v rámci ktorých sú na Stredný východ vyvážané drony s inzerovanými schopnosťami smrtiacich autonómnych zbraňových systémov (LAWs – budeme o nich ešte pojednávať).
Napr. vrtuľníkový dron Blowfish A3 od čínskej firmy Ziyan, je vybavený ľahkým guľometom a je schopný operovať na komplikovaných bojových misiách, realizovať prieskumné úlohy a tiež vykonávať vysoko presné likvidačné údery.
TUCKER, P. *SecDef: China Is Exporting Killer Robots to the Mideast*. [on-line]. [cit. 6. decembra 2021]. Dostupné na internete: <https://www.defenseone.com/technology/2019/11/secdef-china-exporting-killer-robots-mideast/161100/>

- vedenie vojny v kybernetickom priestore

2.7.1. Vojenské spravodajstvo

Využitie systémov umelej inteligencie v oblasti vojenského spravodajstva má na rozdiel od využitia technológií AI v spravodajských službách iných zložiek štátu užší záber činnosti, ktorý mimo autokratických a diktátorských režimov nemá až taký dopad na jednotlivcov a spoločnosť. Na druhej strane však z dôvodu bezpečnosti štátu má oveľa menšie legislatívne obmedzenia (a požíva spoločenský konsenzus) než v civilnom spravodajskom využití. Toto sa ešte stupňuje v režime ohrozenia alebo vojny, keď väčšina regulačných mechanizmov padá, resp. sú nahradené inými.

Čo sme o technológiách AI v spravodajských službách uvádzali v minulej kapitole, môžeme vztiahnuť aj na vojenské systémy. I v oblasti vojenského spravodajstva vzhľadom na limity, riziká a možnosti zneužitia súčasných systémov AI platí požiadavka veľkej opatrnosti a striktnej zákonnosti i potreba dôsledného dohľadu na to určených armádnych štruktúr a mechanizmov, aby prostriedky umelej inteligencie neboli zneužité, resp. sa vymkli kontrole. Vzhľadom na obmedzenejšie možnosti demokratického dohľadu a kontrolných mechanizmov – osobitne v čase ohrozenia alebo vojny – to nemusí byť jednoduché zabezpečiť. Hrozbou je i využitie v autokratických režimoch a diktatúrach s prienikom armádnej a občianskej roviny spoločnosti.

2.7.2. Modelovanie technológií, konfliktov a operácií

Modelovanie technológií, konfliktov a operácií je zaujímavou oblasťou využitia systémov AI, ktorá – ak sa bavíme o technológiách – osobitne vo vede slávi nebývalé úspechy. Modelovanie zložitých dejov na kvantovej úrovni, štiepnych a fúznych reakcií, makrokozmických astrofyzikálnych dejov,[299] látok v rámci materiálovej fyziky a chémie, biologických a biochemických reakcií – to všetko je nielen príkladom pre využitie vo vojenskom sektore, ale v mnohých oblastiach i priamo súčasťou vojenského výskumu.

299 Ako sme uvádzali v poznámkovom aparáte k úvodnej kapitole – AI prvý raz simulovala vesmír: rýchlo, presne a nikto nevie ako.
SUMMER, *The first AI universe sim is fast and accurate—and its creators don't know how it works*. [online]. [cit. 6. augusta 2020].
Dostupné na internete: <https://phys.org/pdf480780725.pdf>
Odvtedy vzniklo a stále vzniká viacero ďalších simulácií vesmíru, resp. jeho častí, pri ktorých sa algoritmy AI neustále zlepšujú a simulácie sú stále presnejšie a detailnejšie.

Na nasadenie systémov AI v rámci modelovania vojenských technológií môžeme mať dva pohľady. V tom prvom ide o pozitívnu vec, keď sa prostredníctvom modelovania znižuje nebezpečenstvo prameniace z iných foriem výskumu. Ide napríklad o vývoj nových jadrových zbraní, ktoré nevybuchujú na jadrových polygónoch, ale len vo vizualizáciách superpočítačov[300], modelovanie nových prvkov riadených striel, stíhacích lietadiel piatej a šiestej generácie[301], odolných materiálov a pod.

Druhý pohľad je skôr varovný. Schopnosť realisticky modelovať a simulovať môže byť lákadlom pre vývoj zakázaných vojenských prostriedkov, napríklad biologických a chemických zbraní...[302]

Nech sa však na problematiku modelovania a simulovania vojenských technológií pozeráme akokoľvek, vždy treba mať na pamäti limity a riziká uvedené v kapitolách 2.1. až 2.4. Ak by sme ich nebrali do úvahy, akýkoľvek presun technológie alebo zbraňového systému z virtuálnej simulácie do reálneho sveta môže byť ak nie tragédiou, tak minimálne trpkým sklamaním.

Zaujímavým využitím algoritmov umelej inteligencie je modelovanie takticko-operačných stratégií a postupov. Ak je pre tento účel využitý kvalitne natrénovaný systém AI a je k dispozícii dostatok taktických a operačných dát, môže ísť o nezanedbateľný vklad k optimalizácii armádneho nasadenia. Ovocím môže byť nielen zníženie strát na vojakoch a technike v rámci správnej nasimulovanej operácie, ale – ak je pre to záujem a vôľa – i zníženie kolaterálnych strát, t.j. obetí z radov civilného obyvateľstva, životne dôležitej infraštruktúry a pod. Toto ovocie sa však môže stať trpkým, ak by obrovská technologická prevaha vo využití systémov AI bola zneužitá na napadnutie iných štátov alebo upevnenie

300 Ako príklad môžeme uviesť národné laboratóriá USA pre výskum a vývoj jadrových zbraní v Los Alamos. Aspoň jeden z ich superpočítačov sa už tradične umiestňuje v prvej desiatke najvýkonnejších superpočítačov sveta.
DOE/NNSA/LANL/SNL [on-line]. [cit. 3. marca 2022].
Dostupné na internete: <https://www.top500.org/site/50334/>

301 Tieto stroje sú samy o sebe AI enabled, t.j. intenzívne využívajú svoje vlastné lokálne systémy AI a participujú aj na externých systémoch AI, napr. v rámci skupinového riadenia bojovej operácie.

302 Na základe jednoduchých zmien v nastaveniach systém AI MegaSyn navrhol miesto nových liekov niekoľko desiatok vražedných chemických bojových látok.
URBINA, F., LENTZOS, F., INVERNIZZI, C. et al. *Dual use of artificial-intelligence-powered drug discovery*. In: *Nat Mach Intell*. [on-line]. 2022, 4, s. 189–191. [cit. 22. marca 2022].
Dostupné na internete: <https://doi.org/10.1038/s42256-022-00465-9>

diktatúry.

S jedlom rastie chuť, prečo potom nepristúpiť od modelovania operácií k modelovaniu celých konfliktov? Myslíme si, že v rámci súčasných možností systémov umelej inteligencie a od toho sa odvíjajúcich problémov, ktoré sme diskutovali v závere kapitoly 2.1.2, reálne modelovanie celých konfliktov nie je na programe dňa. Možno parciálne modelovanie jednotlivých krokov prebiehajúceho konfliktu môže byť veleniu vojenských operácií nápomocné, no viac skutočne reálnych a spoľahlivých výsledkov sa za súčasného stavu rozvoja systémov AI asi dosiahnuť nedá. Rizikom by sa teda mohlo stať, že by armádne vedenie, resp. vládni činitelia uverili predikciám, ktoré by potencionálne modelovanie mohlo prinášať. I keď túto dôveru môžeme považovať za nereálnu až iracionálnu, niektoré varovania uvedené v kapitole 2.5. a technologické mýty[303], ktoré informačnú spoločnosť sprevádzajú, by mohli k tejto klamlivej dôvere viesť. A bola by to veľká tragédia, ak by ovocím takejto dôvery bolo rozpútanie vojnového konfliktu...

2.7.3. Podpora pre velenie

Rozvíjaním našej rozpravy o modelovaní takticko-operačných stratégií a postupov i simulovaní konfliktov sme priamo prešli do oblasti podpory pre velenie, pričom v niektorých aspektoch sa delenie na modelovanie a podporu prakticky stiera a naviac sa kombinuje s výsledkami vojenského spravodajstva z oblasti analýzy a predikcie.[304]

Podpora pre velenie však prináša aj ďalšie možnosti využitia systémov umelej inteligencie. Ide napríklad o modelovanie logistických potrieb a trás zásobovania, manažment telekomunikačného zabezpečenia velenia, presné meteorologické predpovede a mnoho iných služieb podpory. Niektoré z nich môžeme považovať za natoľko zabehnuté

303 Ide o sociologický fenomén prehnanej dôvery vo funkčnosť, spoľahlivosť a presnosť systémov IKT a AI, ktorá sa radí medzi tzv. mýty informačnej spoločnosti a prejavuje sa ako rizikový faktor napríklad v oblasti kybernetickej bezpečnosti.
Konkrétnejšie sa problematike dôvery voči systémom AI venujeme o niekoľko kapitol ďalej.

304 Podľa Petra Laytona má najnovšia generácia AI vplyv v piatich hlavných oblastiach vrátane identifikácie, zoskupovania, generovania, predpovedania a plánovania. Ľudia môžu tieto činnosti vykonávať, ale umelá inteligencia dokáže tieto úlohy vykonávať efektívnejšie a oveľa rýchlejšie.
LAYTON, P. *Fighting Artificial Intelligence Battle: Operational Concept for Future AI-Enabled Wars*. [on-line]. Australian Defence College, 2021. [cit. 10. októbra 2022].
Dostupné na internete: <https://tasdcrc.com.au/wp-content/uploads/2021/02/JSPS_4.pdf>

a overené, že s vážnejšími zlyhaniami ani nerátame, resp. zlyhania sú zakomponované do riadenia rizík.[305]

Iné však v sebe môžu skrývať riziká špecifické pre algoritmy AI. Napríklad zlyhanie zásobovania počas prebiehajúceho konfliktu alebo humanitárnej operácie, ktorého dôvodom je nesprávne natrénovaný systém AI riadiaci logistické procesy. Ak totiž máme systém trénovaný na dátach možno aj veľmi úspešných vojnových operácií, avšak riadených iným veliteľským štýlom, než aký používa aktuálne velenie (iné vedenie operácií môže znamenať inú [s]potrebu materiálu, pohonných hmôt, podporného personálu, ženijnej podpory a pod.), generované požiadavky na logistiku môžu byť nesprávne.

2.7.4. Trenažéry, simulátory a výcvik

Tréningové, operačné i taktické dáta (aj) zo sledovania, analytika i predikcia, modelovanie a simulácie – to všetko sa spája v moderných trenažéroch, simulátoroch a výcvikových zariadeniach využívajúcich systémy AI.

Ide o technológie, ktoré vďaka algoritmom umelej inteligencie dávno presahujú schopnosti bežných tréningových nástrojov a pomôcok, dokážuc hodnoverne simulovať reálne bojové nasadenie a v rámci výcviku aktívne vystupovať ako skutočne schopní protivníci najrozličnejšieho druhu.

Výhodou nasadenia simulátorov vybavených technológiami AI je zabezpečenie kvalitnej odbornej i taktickej prípravy vojenského personálu a možnosť dlhodobo ho udržať v dobrej kondícii. Naviac bez zbytočných rizík a extrémnych nákladov pre zabezpečenie tak kvalitnej prípravy.

Rizikovým faktorom je psychologický aspekt vytvorenia si podvedomého chápania boja ako virtuálnej reality. **Dôsledkom môže byť strata citlivosti pre ohrozenie ľudských životov a reálnych hodnôt pri nasadení v skutočnom boji**. Tento rizikový faktor je o to záludnejší, že sa „pod kožu“ dostáva postupne a podvedome, mnohokrát bez varovnej reakcie svedomia.

Ako sme spomenuli, trenažéry a simulátory bojových operácií na báze umelej inteligencie vedia vystupovať ako zdatní súperi vo virtuálnych bojových situáciách. Logickým

305 Bez riadneho manažmentu rizík nie je možné hovoriť o spoľahlivom a bezpečnom nasadení prostriedkov IKT a AI v životne dôležitých oblastiach.
Risk management [on-line]. [cit. 3. marca 2022].
Dostupné na internete: <https://en.wikipedia.org/wiki/Risk_management>

dôsledkom je tak ich možnosť nasadenia v reálnych zbraňových systémoch...

2.7.5. Autonómne zbraňové systémy

Jednou z najdiskutovanejších oblastí využitia systémov umelej inteligencie v armádnej sfére sú autonómne zbraňové systémy. Ide o armádne vybavenie, u ktorého sa snúbi pokročilé technologické a zbraňové vybavenie s adaptabilnosťou a autonómnosťou systémov umelej inteligencie.

Autonómne zbraňové systémy by sme v zásade mohli rozdeliť na smrtiace a ostatné, zahŕňajúce široké portfólio bojovej techniky od prieskumných operácií, cez komunikačnú a logistickú podporu až napríklad po ženijné nasadenie.[306]

Z pohľadu nasadenia systémov AI vo vojenskej oblasti však zásadnú roľu hrajú smrtiace autonómne zbraňové systémy[307], schopné na základe činnosti algoritmov umelej inteligencie samostatne vyhľadávať, identifikovať a zasahovať ciele. Ide o systémy pokrývajúce najrozličnejšie zbraňové technológie a schopné operovať vo vzduchu, na zemi, vo vode, pod vodou alebo vo vesmíre.[308]

Náš záujem v kontexte etických výziev sa prirodzene zameriava práve na tieto smrtiace autonómne zbraňové systémy...

Podľa niektorých vojenských expertov autonómne zbraňové systémy nielenže prinášajú významné strategické a taktické výhody na bojisku, ale sú aj z morálnych dôvodov vhodnejšie v porovnaní s nasadením ľudského personálu. Kritici naopak zastávajú názor, že tieto zbrane by sa mali obmedziť, ak nie úplne zakázať, a to z rôznych morálnych a právnych dôvodov.[309]

Protagonisti LAWs teda obhajujú ich využitie v dvoch rovinách – z dôvodu vojenských

---

306 Bojová technika z portfólia ostatných autonómnych zbraňových systémov sa vo viacerých oblastiach priamo či nepriamo prekrýva s inými oblasťami systémov AI využívaných v armáde.

307 **Lethal autonomous weapon – používa sa skratka LAWs**, prípadne len AWS, keďže použitie skratky LAWS môže byť v anglickom texte mätúce.

308 *Lethal autonomous weapon* [on-line]. [cit. 5. marca 2022].
Dostupné na internete: <https://en.wikipedia.org/wiki/Lethal_autonomous_weapon>

309 ETZIONI, A., ETZIONI, O. *Pros and Cons of Autonomous Weapons Systems* [on-line]. [cit. 5. marca 2022].
Dostupné na internete: <https://www.armyupress.army.mil/Journals/Military-Review/English-Edition-Archives/May-June-2017/Pros-and-Cons-of-Autonomous-Weapons-Systems/>

výhod a z dôvodu morálnosti ich nasadenia.

Medzi vojenské výhody radia ponajprv ich multiplikačný efekt, čo znamená, že na vykonanie bojovej misie je potrebný menší počet vojakov, pričom účinnosť každého z nich je vyššia. Ďalším benefitom je rozšírenie bojiska, keďže LAWs sú schopné preniknúť do oblastí, ktoré by boli bez ich nasadenia inak nedostupné. A v neposlednom rade je dôležitou výhodou i nahradenie ľudského faktora v nebezpečných misiách a tým aj minimalizovanie strát na životoch, čo má ďalekosiahle dôsledky – od psychologických v rámci armády, cez spoločenské (verejná mienka a podpora je dôležitým faktorom pre bojové operácie) až po materiálne a odborné.[310]

Ďalšie zaujímavé výhody uvádza aj cestovná mapa nasadenia bezpilotných systémov pre roky 2007-2032: robotické systémy sú vhodnejšie ako ľudia pre nasadenie na „nudné, špinavé alebo nebezpečné“ misie. Medzi nudné misie patria dlho-trvajúce (monotónne) operácie. Ako príklad špinavej misie sa uvádza vystavenie personálu potenciálne škodlivému rádioaktívnemu materiálu. A príkladom nebezpečnej misie môže byť zneškodňovanie výbušnín, či boj pod intenzívnou paľbou protivníka.[311]

Medzi ďalšie výhody patrí potenciál operovať v rýchlejšom tempe, ako je možné dosiahnuť u ľudí a schopnosť smrteľne zasiahnuť aj vtedy, keď je prerušené komunikačné spojenie.[312] V dlhodobom horizonte ide aj o nemalé úspory na personále – jeho vzdelaní,

---

310 Por. MARCHANT, G. E. et al. *International Governance of Autonomous Military Robots*. In: *Columbia Science and Technology Law Review.* [on-line]. 2011, 12, s. 272–76. [cit. 27. marca 2017]. Dostupné na internete: <http://stlr.org/download/volumes/volume12/marchant.pdf>
Ohľadom materiálnych a odborných dôsledkov minimalizovania strát na životoch je potrebné si uvedomiť, že odborný výcvik, vzdelanie a tréning pre moderné sofistikované zbraňové systémy je časovo i intelektuálne náročný a drahý. Napr. strata bojového lietadla v hodnote desiatok miliónov dolárov bolí, no oveľa viac velenie mrzí, ak pri tom príde aj o schopného pilota, ktorého nenahradia tak ľahko ako havarovaný, či zostrelený stroj.

311 CLAPPER, J. R. et al. *Unmanned Systems Roadmap: 2007-2032*. [on-line]. Washington, DC: Department of Defense [DOD], 2007, s. 19. [cit. 5. marca 2022]. Dostupné na internete: <http://www.globalsecurity.org/intell/library/reports/2007/dod-unmanned-systems-roadmap_2007-2032.pdf>

312 THURNHER, J. S. *Legal Implications of Fully Autonomous Targeting*. In: *Joint Force Quarterly*. [on-line]. 2012, 67, 4, s. 83. [cit. 7. marca 2022]. Dostupné na internete: <http://ndupress.ndu.edu/Portals/68/Documents/jfq/jfq-67/JFQ-67_77-84_Thurnher.pdf>

tréningu a službe.[313]

I keď sme s uvedením všetkých výhod nasadenia LAWs v armáde ešte neskončili, prichádzame k závažnému zisteniu, pre ktoré je a bude obmedzovanie nasadenia smrtiacich systémov veľmi ťažké – **vo viacerých scenároch nasadenia sú zbraňové systémy vybavené technológiami umelej inteligencie lepšie ako ľudia**.

Z pohľadu LAWs to nie je nič nové, ak do kategórie smrtiacich automatických zbraňových systémov radíme nielen systémy AI, ale aj jednoduchšie automatizované zbrane, napr. radarom navádzané systémy CIWS používané na obranu lodí, ktoré sa používajú od 70. rokov 20. storočia. Takéto systémy dokážu autonómne identifikovať a zaútočiť na prichádzajúce rakety, delostreleckú paľbu, lietadlá a hladinové plavidlá podľa kritérií stanovených ľudskou obsluhou.[314]

Moderné technológie umelej inteligencie však schopnosti LAWs povzniesli na úplne novú úroveň, pri ktorej sa pokročilá automatizácia mení na skutočnú adaptabilnosť a autonómnosť.[315]

Ako príklad môžeme uviesť moderné letecké bojové simulátory, pri ktorých – nadviažuc na pasáž o využití systémov AI pri trenažéroch a výcviku – je možné porovnávať schopnosti ľudskej obsluhy sofistikovaných zbraňových systémov a ich umelých náprotivkov. Systém umelej inteligencie ALPHA[316], ktorý pre tento účel vyvinuli na Univerzite v Cincinnati, zdrvujúco porazil v simulovaných bojových operáciách jedného z najlepších inštruktorov stíhacích pilotov v USA, plukovníka Geneho Leea.[317] ALPHA ho

313 David Francis v The Fiscal Times v roku 2013 cituje údaje ministerstva obrany, podľa ktorých „každý vojak v Afganistane stojí Pentagon približne 850 000 dolárov ročne“.
FRANCIS, D. *How a New Army of Robots Can Cut the Defense Budget*. In: *Fiscal Times*. [on-line]. 2013, 2. 4. [cit. 5. marca 2022].
Dostupné na internete: <http://www.thefiscaltimes.com/Articles/2013/04/02/How-a-New-Army-of-Robots-Can-Cut-the-Defense-Budget>

314 V lodných systémoch ide napr. o americké systémy CIWS Phalanx, pri tankových napr. ruský systém Arena, izraelský Trophy a nemecký AMAP-ADS.

315 Ide o základné vlastnosti každého systému umelej inteligencie, ktoré sme uviedli v kapitole 1.3.

316 Ide o fuzzy orientovaný systém umelej inteligencie s genetickými algoritmami (genetic-fuzzy systems).

317 Inštruktor, ktorý vyškolil tisíce pilotov US Air Force, to komentoval slovami: skúsený bojový pilot dokáže bez problémov poraziť väčšinu súčasných bojových umelých inteligencií ako lovnú zver... Pri systéme ALPHA sme lovnou zverou my... Ešte nikdy som sa nestretol s tak agresívnou, vnímavou, dynamickou

dokonale zaskočila – excelentne sa chránila a dokázala okamžite správne reagovať. Výborne predvídala zámery protivníka a pohotovo odpovedala na akékoľvek zmeny letu a odpaľovanie rakiet. Nielenže sa vynikajúco vyhýbala odpáleným raketám, ale vedela podľa potreby aj plynule prechádzať medzi defenzívnymi a ofenzívnymi fázami boja. ALPHA proti plukovníkovi bojovala v simulácii reálnych, niekoľko hodín trvajúcich misií, po ktorých bol Lee totálne vyčerpaný – fyzicky i mentálne na dne. Asi netreba dodávať, že ALPHA bola neustále na koni a nemala problém bojovať ďalej...[318]

Zámerne sme použili príklad simulátora z roku 2016, aby sme si v kontexte neustále pripomínaného veľkého pokroku na poli umelej inteligencie uskutočneného v ostatných rokoch uvedomili, o koľko vyspelejšia je asi súčasná technika LAWs.

Zbraňové systémy vybavené technológiami AI sa vo veľkom presúvajú z oblasti simulátorov do reálnych cvičných a bojových strojov.

Ponajprv išlo a ide o cvičné stroje, ktoré umožňujú využívať výhody umele inteligencie známe z virtuálneho prostredia v reálnom výcviku. Len v USA ide o produkciu viacerých firiem[319], čomu sa snažia úspešne sekundovať výrobcovia z iných krajín.[320]

Akonáhle sa využitie cvičných systémov osvedčilo, prichádzalo na nasadenie v prieskumných a bojových misiách, pričom sa rozdiely medzi prieskumným a bojovým využitím stierajú, takže moderné LAWs vedia byť veľmi univerzálne pri reálnom nasadení.[321]

Na vývoji LAWs v oblasti vojenského letectva je asi najmarkantnejšie vidieť, ako sa moderné automatizované zbraňové systémy približujú špičkovej leteckej technike. Moderné drony sú vybavené podobnými technológiami ako bojové lietadlá generácií 4+

---

a vierohodne bojujúcou umelou inteligenciou, ako je ALPHA.

318 MIHULKA, S. *Letecká bojová umělá inteligence si natřela na chleba taktické experty*. [on-line]. [cit. 5. marca 2022].
Dostupné na internete: <https://www.osel.cz/8903-letecka-bojova-umela-inteligence-si-natrela-na-chleba-takticke-experty.html>

319 Len z najznámejších ide o General Atomics, Boeing alebo Kratos, z tých najnovších so zaujímavými cieľmi vývoja napr. Exosonic alebo Dynetics.

320 V zásade ide o krajiny patriace do štyroch kategórií štátov najviac napredujúcich vo využívaní AI vo vojenskej oblasti, ktoré sme spomínali v úvode tejto kapitoly.

321 Kto by nepoznal klasický dron MQ-9 Reaper od General Atomics – ostrieľaného ~~kozáka~~, pardon kovboja, ktorý už bol nasadený v nespočetných prieskumných a bojových misiách.

a 5 a sú schopné s nimi aj v bojových operáciách priamo spolupracovať.[322]

Zaujímavým prvkom je i kombinácia rôznych technológií v rámci LAWs. Napríklad vývoj moderného lacného[323] bojového dronu Skyborg v máji 2021 ešte nebol hotový, no jeho riadiaci mozog (Autonomy Core System, ACS) poháňaný umelou inteligenciou už bol zrelý na testovanie v reálnej prevádzke. ACS bol preto nainštalovaný na taktický dron Kratos UTAP-22 a poslaný do vzduchu.[324] Zámerne uvádzame aj túto – pre armádny „mainstream“ – nepodstatnú udalosť, keďže ukazuje, do akej miery moderné zbraňové systémy v posledných rokoch pokročili od špeciálne „vypiplaných“ a na mieru vytvorených celkov k univerzálnym riešeniam.

Pokiaľ teda LAWs využívajú algoritmy umelej inteligencie, dokážu byť neuveriteľne technologicky adaptabilné, relatívne lacné a minimálne rovnako nebezpečné ako najmodernejšie zbraňové systémy riadené ľudskou obsluhou.

Čitateľ, pamätajúc na riziká a limity, ktoré sme v kapitolách 2.1. až 2.4. uvádzali, však okamžite spozornie – **ako dokážeme zabezpečiť, že smrtiace autonómne systémy, ktoré sú zámerne dizajnované tak, aby dokázali efektívne pracovať aj pri strate spojenia, budú vedieť správne použiť svoju smrtiacu silu a rešpektovať požadované mantinely, ak je nám známe, že systémy AI robia mnohokrát iné chyby ako ľudia[325] a nevieme ich vytrénovať tak, aby vedeli adekvátne odpovedať na všetky reálne situácie?**

Rozoberajúc autonómne zbraňové systémy sme sa viackrát dotkli ich nasadenia v kooperácii s inými automatizovanými alebo ľudskou obsluhou riadenými systémami.

2.7.6. Skupinové riadenie bojových prostriedkov a autonómnych systémov

Pri skupinovom riadení bojových prostriedkov a autonómnych systémov by sme mohli hovoriť o LAWs na stereoidoch, keďže rôzne výhody uvedené pri ich individuálnom nasadení sa znásobujú. Podstatou skupinového riadenia je spojenie schopností a činnosti

322 Napr. prúdový bojový dron Gambit od General Atomics, Loyal Wingman od Boeingu a pod.

323 Ide o kategóriu tzv. „attritable drones“ – ktoré sú celkom lacné, takže môžu byť nasadené vo vysoko rizikových konfliktoch bez toho, že by vyššie bojové straty ohrozili rozpočet.

324 SZONDY, D. *Skyborg combat drone's "brain" flies for the first time*. [on-line]. [cit. 7. marca 2022]. Dostupné na internete: <https://newatlas.com/military/skyborg-combat-drones-brain-first-time/>

325 Omyly a chyby systémov AI sú iné, než tie ľudské – viď kapitola 2.1. alebo diskusia v kapitole 1.1. o programe Eugene, ktorý ako prvý zvládol Turingov test.

autonómnych zbraňových systémov do jedného harmonického celku, ktorého časti síce plnia svoje vlastné samostatné úlohy, no spoločne sledujú jeden bojový cieľ.[326]

Rozlišujeme dva spôsoby skupinového riadenia a nasadenia autonómnych zbraňových systémov: spoločné nasadenie s bojovými prostriedkami ovládanými človekom a koordinované nasadenie LAWs pod spoločným riadením centrálnym systémom umelej inteligencie, prípadne koordinovaným riadením na základe interakcií jednotlivých autonómnych systémov.

Zbraňové systémy AI spolupracujúce s prostriedkami ovládanými človekom prevažne pracujú v submisívnom režime, pri ktorom bojovú misiu riadi človek. LAWs v tomto režime zvyčajne operujú ako sprevádzajúce a spolupracujúce zbraňové systémy pod velením človeka. Vzhľadom na svoju výbavu systémami AI však dokážu vykonávať parciálne podúlohy bojovej misie nielen autonómne, ale prípadne sú schopné aj prevziať iniciatívu a operovať úplne samostatne.[327]

Mohlo by sa zdať, že nasadenie LAWs je v tomto režime pod kontrolou, no v prípade umelou inteligenciou riadených zbraní to tak vôbec nemusí byť. Súčasťou autonómneho vykonávania parciálnych úloh tiež môže byť použitie smrtiacej sily na základe vyhodnotenia, ktoré vykoná algoritmus AI. Prevzatie iniciatívy sprevádza v zásade to isté riziko, len môže byť umocnené nielen úplne autonómnym rozhodovaním systému AI, ale i dôsledkami aktuálneho vývoja bojovej situácie.[328]

---

326 Americký Defense Science Board používa termín *multiagent coordination* pre okolnosti, v ktorých je úloha rozdelená medzi „viacero robotov, softvérových agentov alebo ľudí“.
Defense Science Board. *Task Force Report: The Role of Autonomy in DoD Systems*. Washington, DC: Office of the Under Secretary of Defense for Acquisition, Technology and Logistics, 2012.

327 Napríklad v dňoch písania tejto kapitoly predstavený plnohodnotný bojový prúdový dron Gambit od General Atomics, ktorý by mal operovať ako sprevádzajúci bojový stroj pilotovaného lietadla a po jeho boku a pod jeho vedením sa zúčastňovať vzdušných bojov. Vzhľadom na svoju výbavu systémami AI však dokáže vykonávať podúlohy bojovej misie autonómne a prípadne aj prevziať iniciatívu a operovať úplne samostatne.
SZONDY, D. *General Atomics' Gambit autonomous combat drone takes the initiative*. [on-line]. [cit. 7. marca 2022].
Dostupné na internete: <https://newatlas.com/military/general-atomics-gambit-combat-drone-air-dominance/>

328 Napríklad zostrelenie veliaceho pilotovaného bojového lietadla v tak špecifickej konkrétnej situácii v rámci mestskej aglomerácie, že následné správne rozhodnutie o použití smrtiacej sily autonómneho

Koordinované nasadenie LAWs pod vedením centrálneho systému AI, resp. spoločného riadenia zapojených systémov AI samozrejme obnáša rovnaké kategórie rizík ako samostatné autonómne zbraňové systémy, len ich dôsledky sú v prípade zlyhania nebezpečne násobené.

Naviac sa vynára viacero nových problémov týkajúcich sa činnosti týchto zbraňových systémov v prípade výpadku spojenia, resp. kolapsu centrálneho systému riadenia. Sú či budú parciálne systémy AI (napríklad u miniatúrnych alebo skôr spomínaných „attritable drones“) tak riešené, aby vedeli bezpečne zvládať aj tieto situácie?

Je koordinované riadenie na základe interakcie jednotlivých systémov AI natoľko robustné, aby prepojené LAWs dokázali správne pracovať aj pri výpadku kritického množstva zapojených automatických systémov?

Vieme dostatočne odhadnúť správanie centrálne riadeného (a zároveň parciálne autonómneho) alebo koordinovane riadeného skupinového automatického zbraňového komplexu vo vyhranených situáciách?

2.7.7. Vedenie vojny v kybernetickom priestore

Last but not least, treba spomenúť využitie systémov umelej inteligencie v rámci vedenia vojny v kybernetickom priestore.[329] Kybernetický priestor v moderných doktrínach predstavuje ďalší z rozmerov vedenia vojny, ktorý sa v poslednom desaťročí stáva tak významným, že si vo viacerých armádach vydobyl svoje pevné postavenie v rámci samostatnej armádnej zložky. Mnohí experti ho nazývajú bojovým priestorom 21. storočia.[330] Kybernetický priestor zároveň predstavuje novú operačnú doménu, ktorej charakter a prepojenosť s ostatnými operačnými doménami – zem, voda, vzduch a vesmír – vytvára priestor pre synergiu a efektívne použitie vojenskej sily, ekonomizáciu vynakladaných prostriedkov a dosahovanie zámerov velenia v reálnom čase.[331]

dronu v zastavanej oblasti je nad natrénovaný rámec jeho systému AI.

329 Kybernetické zbrane – cyberwarfare.

330 Ak v moderných dejinách vojny bolo 19. storočie arénou pozemných jednotiek a námorníctva a 20. storočie ovládol strategický význam letectva, virtuálny priestor vytvorený celosvetovou digitálnou interkonektivitou má byť priestorom kybernetického boja.
GIBNEY, A. *Zero Days*. [filmový dokument]. [cit. 10. marca 2022].
Dostupné na internete: <http://www.zerodaysfilm.com/>

331 Z predkladaného Návrhu Stratégie kybernetickej obrany Slovenskej republiky.
Stratégia kybernetickej obrany Slovenskej republiky. [on-line]. Ministerstvo obrany SR, Vojenské

Fenomén kybernetického vedenia vojny stavia na dvoch pilieroch – využití informačných a komunikačných technológií v spravodajstve a pokroku v oblasti kybernetickej bezpečnosti, pričom treba zdôrazniť, že pomyselné hranice medzi kybernetickým vedením vojny a spravodajstvom, resp. kybernetickou bezpečnosťou sú veľmi vágne a mnohé kybernetické operácie je ťažké presne zaradiť.

Spravodajský pilier sme načrtli v predchádzajúcej kapitole. Pilier kybernetickej bezpečnosti[332] je vhodné trošku konkretizovať, keďže pre účely vojenského využitia kybernetická bezpečnosť ide ruka v ruke s využitím metód a nástrojov kybernetickej kriminality.

Bezpečnosť v oblasti IKT prešla v posledných desaťročiach búrlivým vývojom – od praktickej neexistencie, či úplného marginalizovania, cez riešenie problémov, ktoré vznikali pri reálnej prevádzke a potrebe chrániť systémy pred dôsledkami rozvíjajúcej sa počítačovej kriminality, až po dnešný stav, v ktorom ide o dosť rozsiahly vedecký obor, jeden z podstatných a nutných rozmerov dizajnu moderných informačných technológií a vytváranie jednotiek kybernetického boja u armád nielen najsilnejších štátov sveta. Zároveň z pohľadu informačnej kriminality ide v súčasnosti už o najvýnosnejší spôsob kriminality, predstihujúc tak drogovú kriminalitu a obchodovanie s ľuďmi.[333]

---

spravodajstvo, s. 3. [cit. 17. marca 2022].
Dostupné na internete: <https://www.slov-lex.sk/legislativne-procesy/-/SK/dokumenty/LP-2022-128>

332 Používame termín kybernetická a nie informačná bezpečnosť, keďže kybernetická bezpečnosť je podmnožinou informačnej bezpečnosti zahŕňajúcej aj oblasti, ktoré nie sú predmetom našej diskusie.

333 Kybernetická kriminalita sa v rokoch 2015-2016 stala najvýnosnejšou oblasťou kriminality.
Cybercrime Is Now More Profitable Than The Drug Trade [on-line]. [cit. 10. marca 2022].
Dostupné na internete: <https://www.tripwire.com/state-of-security/regulatory-compliance/pci/cybercrime-is-now-more-profitable-than-the-drug-trade/>
Pokročilý ransomware je v súčasnosti najvýnosnejším obchodným modelom kybernetickej kriminality.
Why Advanced Ransomware Is Cybercrime's Most Profitable Business Model [on-line]. [cit. 10. marca 2022].
Dostupné na internete: <https://blog.knowbe4.com/why-advanced-ransomware-is-cybercrimes-most-profitable-business-model>
ŠANTAVÝ, *Analýza virtuálneho sveta v kontexte náboženskej formácie, pôsobenia a života Cirkvi,* [on-line], s. 129. [cit. 10. marca 2022].
Dostupné na internete: <https://peter.santavy.cloud/data/uploads/docs/analyza_virtualneho_sveta_v_kontexte_nabozenskej_formacie_posobenia_a_zivota_cirkvi.pdf>

Kybernetická kriminalita stavia na realite chýb v prostriedkoch IKT[334], ktoré umožňujú vniknutie do systémov, únik informácií a dokonca prevzatie vlády nad systémom.

S nástupom masívneho využívania prostriedkov IKT v mnohých oblastiach života a fungovania spoločnosti prichádzalo k postupnej kriminalizácii zneužívania chýb.

V rámci prebiehajúceho procesu prechodu k informačnej a znalostnej spoločnosti, ktorá je bytostne spätá s fungujúcimi prostriedkami IKT, kybernetická bezpečnosť, resp. jej temná kriminálna stránka nadobúdajú strategickú dôležitosť a stávajú sa predmetom útočných i obranných aktivít v čase vojnových konfliktov.

K charakteristickým črtám prvkov kybernetického vedenia vojny tiež patrí ich zaradenie k tzv. prostriedkom vojen štvrtej generácie[335], ktoré sú popisované stieraním hraníc medzi vojnou a politikou[336], medzi armádou a civilným obyvateľstvom, decentralizovaným vedením vojny, guerilovou taktikou a prvkami terorizmu, dezinformačným pôsobením a propagandou, útokom na kultúru a psychologickými metódami na oslabenie protivníka.[337]

S rastúcou účinnosťou kybernetických prostriedkov, potencionálnym efektívnym dopadom na nepriateľa a výhodami v rámci vedenia vojenských operácií sa vo viacerých armádach kybernetické jednotky postupne etablujú ako samostatná armádna zložka.

V oblasti kybernetickej bezpečnosti, v ktorej sa autor profesionálne pohybuje, osobitne v posledných desaťročiach badať nielen uskutočnenie prvých úspešných a sofistikovaných útokov vedených priamo a len prostriedkami kybernetického vedenia vojny, ale aj neustále ataky, ktoré do tejto oblasti patria. Ide napríklad o testovanie štruktúr IKT niektorých štátov (v posledných čase realizované napr. v špecializovaných DDoS útokoch na štátnu

334 Ide o chyby v hardvéri, základnom softvérovom vybavení (firmware, operačné systémy), aplikačnom softvéri, komunikačných technológiách (chyby šifrovania a pod.).

335 *Fourth-generation warfare*. [on-line]. [cit. 10. marca 2022].
Dostupné na internete: <https://en.wikipedia.org/wiki/Fourth-generation_warfare>

336 Tiež treba podotknúť, že aktérmi nemusia byť len štáty, resp. štátne zoskupenia. Môže ísť nielen o zástupné organizácie niektorých štátov, ale aj o akékoľvek iné mimovládne činitele.

337 Zaujímavým konkrétnym príkladom komplexnosti štvrtej generácie je štúdia prestížneho amerického think-tanku RAND Corporation o možnostiach destabilizácie a ekonomického vyčerpania Ruska. DOBBINS, J., COHEN, R. S., CHANDLER, N. et al. *Overextending and Unbalancing Russia: Assessing the Impact of Cost-Imposing Options.* [on-line]. Santa Monica, CA: RAND Corporation, 2019. [cit. 18. marca 2022].
Dostupné na internete: <https://www.rand.org/pubs/research_briefs/RB10014.html>

infraštruktúru Estónska a Česka), útoky na komplexné siete internetu vecí, ktoré riadia energetické sústavy, tepelné rozvody, diaľkové distribučné siete fosílnych palív[338], a pod.

Priam exemplárny vzhľad do problematiky kybernetického vedenia vojny ponúka Stuxnet[339], ktorý bol v roku 2007[340] nasadený voči jadrovým zariadeniam v Iráne a ktorý predstavoval v zásade prvý, zároveň najväčší a najsofistikovanejší prípad štátom, resp. viacerými štátmi vykonaného kybernetického útoku.

Stuxnet bol po prekročení svojich operačných hraníc a rozšírení do viacerých častí sveta objavený bieloruskou bezpečnostnou službou VirusBlokAda v roku 2010 a následne analyzovaný špecialistami firiem Kaspersky a Symantec, ktoré sa zameriavajú na kybernetickú bezpečnosť.

Stuxnet bol vyvinutý pre útoky na systémy SCADA, prostredníctvom ktorých boli ovládané centrifúgy (odstredivky) používané na separáciu jadrového materiálu.

Po stránke technologickej Stuxnet spôsobil v bezpečnostnej komunite šok, keďže využíval ukradnuté originálne bezpečnostné certifikáty, dokázal atakovať veľmi špecifický hardvér[341], využíval štyri softvérové chyby nultého dňa (to najlepšie, o čom drvivá väčšina kybernetickej scény ani nevedela), softvérový útok vykonával veľmi cielene a nezaútočil na čokoľvek a akokoľvek, dokázal rozlíšiť dôležitosť cieľa a podľa toho útočiť a i keď infikoval zariadenia prakticky po celom svete, vedel sa aktivovať len v konkrétnych destináciách v Iráne. Stuxnet využíval aj vynikajúce krytie, či už na tú dobu pokročilé techniky skrývania v systéme, alebo priamo počas útoku, keď centrifúgy privádzal k seba deštrukcii, obsluhu totálne zmiatol podávaním klamlivých informácií na ovládacích

338 Tvrdý útok na Colonial Pipeline, najväčšiu distribučnú sieť fosílnych palív v USA, za ktorým stála hackerská skupina DarkSide a ktorý prebehol začiatkom mája 2021.
Colonial Pipeline offline due to ransomware attack. [on-line]. [cit. 13. marca 2022].
Dostupné na internete: <https://www.fortiguard.com/outbreak-alert/darkside>

339 *Stuxnet*. [on-line]. [cit. 10. marca 2022].
Dostupné na internete: <https://en.wikipedia.org/wiki/Stuxnet>

340 Stuxnet bol vyvinutý pravdepodobne v roku 2005, jeho nasadenie s infikovaním iránskych SCADA systémov sa uskutočnilo v roku 2007. V roku 2009 bol Stuxnet zaznamenaný aj v iných zariadeniach a v roku 2010 sa začal šíriť vo väčšom meradle.
NSA – minimálne po odhalení – Stuxnet uvádzala pod názvom Olympijské hry (Olympic Games).

341 V rámci SCADA išlo o PLC (programovateľné logické radiče) od Siemensu, konkrétne špecifické dve verzie frekvenčných meničov, použitých v centrifúgach jadrových zariadení v Iráne.

paneloch. Nikto nevedel, čo sa vlastne deje.

Naviac dizajn a architektúra (i použitý vektor útoku) Stuxnetu nie sú špecifické len pre špecializované verzie PLC v centrifúgach, ale môžu byť prispôsobené ako platforma pre útoky na široké spektrum moderných systémov SCADA a PLC (napr. v továrenských montážnych linkách alebo elektrárňach, rozvodných sieťach, plynovodoch a ropovodoch, atď.), z ktorých väčšina sa nachádza v Európe, Japonsku a Spojených štátoch...

Stuxnet vznikal v čase rozvíjajúcich sa nukleárnych ambícií Iránu s potvrdeným spravodajským materiálom o obohacovaní uránu vo veľkom. Podľa zdroja z NSA išlo o politické rozhodnutie na medzinárodnej a medzirezortnej úrovni so zámerom vyvinúť a nasadiť nový druh zbraní. Aktívne sa na vývoji podieľali USA (CIA, NSA) a Izrael (Mosad), pričom spravodajsky celú operáciu podporila aj britská GCHQ.[342]

NSA síce nemá právomoc zasahovať, no už skôr bol vytvorený US Cyber Commands[343] – vojenské oddelenie, ktoré bolo zodpovedné za národnú bezpečnosť v kybernetickom priestore a majúce spoločné velenie s NSA.[344]

Samotný Mosad mal už v tom čase niečo podobné – kybernetickú jednotku vojenskej rozviedky 8200, špecializujúcu sa na elektronický a kybernetický boj.

Celú operáciu viedla CIA, pričom Mosad bol ten, kto prakticky určoval tempo.

US Cyber Commands spolu s jednotkou 8200 boli zodpovedné za vytvorenie kódu, testovanie prebehlo v testovacom stredisku NSA Oak Ridge na odstredivkách, ktoré odovzdala Líbya. Výsledok testovania bol predložený do Bieleho domu na schválenie operácie.[345]

---

342 Koniec vlády prezidenta Busha bol sprevádzaný **zvyšujúcim sa tlakom na financovanie útočných kybernetických zbraní**. V rámci ministerstva obrany, ktoré v tom čase viedol Robert Gates, prevládala neistota, kam zaradiť kybernetické zbrane, preto bol Stuxnet priradený nie armáde, ale spravodajským službám. To bol dôvod pre účasť CIA, NSA, Mosadu a GCHQ. Podľa niektorých zdrojov bol do zbierania informácií a dodania USB kľúča so Stuxnetom zaangažovaná aj holandská spravodajská agentúra AIVD.

343 US Cyber Commands vzniklo v rámci operácie Buckshot Yankee, t.j. objavenia a riešenia prieniku do tajných sietí Ministerstva obrany USA (CENTCOM) v r. 2008.

344 Na operácii sa podieľalo i oddelenie TAO-S321, hackerské oddelenie, ktoré ako jediné malo právomoc kompromitovať (hackovať) a vstupovať do cudzích systémov.

345 Vzhľadom na schválenie úradujúceho prezidenta Georga W. Busha bolo do kódu Stuxnetu implementované časové obmedzenie jeho aktívneho pôsobenia (niekoľko dní pred inauguráciou nového prezidenta). Po nástupe Baracka Obamu bolo treba nové schválenie, ktoré prišlo hneď v prvom roku

Ovocie je známe, keďže úspešnému útoku Stuxnetu sa pripisuje niekoľkoročné zdržanie iránskeho jadrového programu spôsobením fyzickej deštrukcie približne pätiny iránskych centrifúg (nehovoriac o následnej anabáze s odstavením, analýzou a preverovaním, odvírením a postupnou obnovou funkčnosti ostatných zariadení).

Na začiatku nášho predstavenia Stuxnetu sme uviedli, že predstavoval v zásade prvý, zároveň najväčší a najsofistikovanejší prípad štátom, resp. viacerými štátmi vykonaného kybernetického útoku. Išlo o dôležitý míľnik, ktorý ukázal vysokú efektívnosť využitia kybernetických nástrojov pre spravodajské služby a armádu.[346]

Išlo o štátom riadený útok, na ktorom participovali viaceré štáty a ich rôzne inštitúcie. Stuxnetom tak nastal na tejto úrovni posun od koordinovaného získavania kvalitných spravodajských informácií a obranných operácií k odhodlaniu vykonať útočné aktivity proti inému štátu.

V rámci vojnového použitia ide o ďalšiu z revolučných zmien, ktoré menia pohľad na bojové operácie – tie sú vykonávané elektronicky, pritom však s reálnym a ťažkým dopadom na protivníka, môžu byť realizované na veľké vzdialenosti a prakticky v neobmedzenom rozsahu, môžu mať extrémne rýchly či naopak nebadaný priebeh a minimálne znaky identifikujúce útočiace či bojujúce strany.

Stuxnet bol zároveň súčasťou zmeny v činnosti spravodajských služieb s dopadom na armádne zložky, keď miesto obranných aktivít v kybernetickom priestore sa začal dávať dôraz na útočné aktivity voči iným štátom, resp. aktérom konfliktu. Ruka v ruke s tým, podobne ako v iných oblastiach vojenského nasadenia elektronických systémov, badať stieranie rozdielov medzi obrannými a útočnými aktivitami – tie isté kybernetické zbrane môžu sledovať či brániť, ale i útočiť.

jeho vlády (2009).

346 Zo známych jednotiek kybernetického boja môžeme okrem uvedených v kauze Stuxnet spomenúť aspoň niektoré ďalšie:
Napr. čínska jednotka č. 61419 spadá pod špeciálnu zložku Sily strategickej podpory, ktorá má okrem iného na starosti kybernetické vedenie vojny.
The People's Liberation Army Strategic Support Force. [on-line]. [cit. 13. marca 2022].
Dostupné na internete: <https://jamestown.org/program/the-peoples-liberation-army-strategic-support-force-update-2019/>
Alebo hackerská a špionážna skupina APT28, známa aj pod označením Fancy Bear, ktoré je podľa nórskej tajnej služby (PST) napojená na ruskú vojenskú rozviedku GRU, konkrétne na jej 85. centrum špeciálnych operácií (GTsSS).

V poznámkach pod čiarou sme vyjadrili, že už koniec vlády prezidenta Georga W. Busha bol sprevádzaný zvyšujúcim sa tlakom na financovanie útočných kybernetických zbraní. Vzhľadom na intenzívne využívanie prostriedkov IKT v tzv. kritickej infraštruktúre USA prinieslo následné obdobie Obamovej administratívy oficiálny dôraz na kybernetickú obranu, pričom sa zdalo, že útočné kybernetické zbrane boli akoby tabu. Z vtedajších rozpočtových údajov však jasne vyplýva, že väčšina kybernetických vojenských výdavkov predsa len išla na útočné zbrane.[347] Toto tvrdenie ostatne podporuje už spomenuté opätovné schválenie Stuxnetu v roku 2009 a napríklad aj neskôr zverejnené štatistiky využívania dronov a všeobecne LAWs v konfliktoch na Blízkom východe.

Stuxnet sa tak stal priekopníkom vývoja nového druhu zbraní a spôsobu ich nasadenia. Bol nie evolúciou, ale revolúciou vo svete kybernetických hrozieb. A i keď priamo v kóde Stuxnetu by sme márne hľadali súčasné technológie umelej inteligencie, išlo o samostatne pracujúci kód, vykonávajúci útočné operácie bez priameho riadenia vzdialeným operátorom.

Jedným z dôsledkov úspešného nasadenia Stuxnetu boli i prvé počiny na poli zaangažovania umelej inteligencie v kybernetických zbraňových systémoch, a to v oblasti obranných systémov. Išlo totiž o hrozby, ktoré neboli dopredu presne definované – nebezpečné zraniteľnosti nultého dňa (t.j. dovtedy neznáme), sofistikovaný spôsob šírenia a maskovania kódu, novátorský prístup útoku a pod. Pre moderné systémy AI a algoritmy hlbokého učenia to bola ponuka, ktorá sa neodmieta a preto logicky prišlo k adaptácii týchto technológií aj v oblasti vojenských kybernetických systémov.[348]

V súčasnosti sú technológie umelej inteligencie plne integrované v obranných vojenských systémoch, pričom ich význam rastie ruka v ruke s ich technologickým rozvojom. Postupne sa prichádza aj k výraznejšiemu využitiu v scenároch útočného nasadenia.

Tu však môže byť problém. Už Stuxnet ukázal, že **ani tak sofistikovaná zbraň nie je úplne pod kontrolou, keďže sa rozšíril aj do iných systémov, než na ktoré bol zameraný a zaútočil aj na iné služby a organizácie**.

Aby toho nebolo málo, do hry vstupovali aj neštátni aktéri, napr. Equation Group[349]

347 Výdavky boli skryté pod rôzne kódové označenia, napr. 10 CN0 (10: armadne, CN0: computer network operations, a pod.).

348 Tieto technológie sú v súčasnosti implementované aj v komerčných bezpečnostných produktoch.

349 *Equation Group*. [on-line]. [cit. 10. marca 2022].

(napojená na už skôr spomínanú jednotku TAO z NSA), ktorí stáli za niektorými zneužitými zraniteľnosťami a spolupodieľali sa na vývoj ďalších kybernetických zbraní (špionážny Flame, súrodenec Stuxnetu Duqu,...) a pod. Z ich produkcie pochádza aj šírenie ďalších verzií Stuxnetu.

Čerešničku na tortu nakoniec pridala hackerská skupina The Shadow Brokers,[350] keď v roku 2017 zverejnila obrovský súbor nástrojov patriacich skupine Equation Group. Spolu so zdrojovými kódmi, ktoré už skôr The Shadow Brokers zverejnili, sa tak prakticky ku každému potencionálnemu aktérovi na poli kybernetických zbraní dostala dostatočná zbrojárska výbava.[351]

V tejto konštelácii však ťažko hovoriť o bezpečnej integrácii algoritmov umelej inteligencie a kybernetických zbraní. Veď ako dokáže ktorýkoľvek aktér napr. zabezpečiť dostatočne veľkú, relevantnú a kvalitnú vzorku trénovacích dát na spoľahlivé vytrénovanie systému? Ako dokáže aktér, ktorý útočný kód len prevzal, prípadne pozmenil, domyslieť, ako sa bude daný systém AI správať v neštandardných situáciách? Podobne môžeme klásť otázky prakticky v každej oblasti rizík, ktoré sme v kapitolách 2.1. až 2.4. predstavili.

A ak technológie umelej inteligencie spojíme s klasickými hackerskými technológiami, ako sú napr. stealth technológie, obfuskovanie kódu, rootkity či polymorfný kód, vieme garantovať bezpečnú funkcionalitu takéhoto systému?

Aj z vlastnej praxe na poli kybernetickej bezpečnosti môžeme povedať, že o kvalite a robustnosti obranných kybernetických systémov využívajúcich algoritmy umelej inteligencie sme sa presvedčili. Avšak **útočné kybernetické systémy využívajúce súčasné technológie umelej inteligencie považujeme za veľmi rizikové a nevhodné pre vojenské nasadenie**. Ich dôsledky by totiž mohli byť katastrofálne.

Kybernetické zbrane sa v súčasnosti naplno integrujú do armádneho portfólia, či už ako priama súčasť a podpora vojenských aktivít[352] alebo ako samostatné zbraňové systémy.

Dostupné na internete: <https://en.wikipedia.org/wiki/Equation_Group>

350 *The Shadow Brokers*. [on-line]. [cit. 10. marca 2022]. Dostupné na internete: <https://en.wikipedia.org/wiki/The_Shadow_Brokers>

351 Táto skutočnosť plne zapadá do nášho zaradenia kybernetických zbraní medzi prostriedky vojen štvrtej generácie.

352 Napríklad začiatok aktuálneho vojnového konfliktu na Ukrajine, ktorý v závere nasledovnej kapitoly tiež spomíname, bol sprevádzaný podpornými útočnými kybernetickými operáciami: po DDoS útoku, ktorý

2.7.8. Ďalšie etické a právne aspekty

Či už ide o využívanie LAWs, dôsledky kybernetického boja alebo zneužitie sily spôsobené nasadením akýchkoľvek iných vojenských systémov AI – k otázkam a výhradám, ktoré sme v rámci ich predstavenia priebežne uvádzali, môžeme pripojiť i ďalšie etické a právne aspekty, zosumarizované v už skôr citovanom dokumente *Pros and Cons of Autonomous Weapons Systems*.[353]

Viaceré z výhrad boli spísané v rámci otvorených listov a výziev požadujúcich či už zakázanie alebo zmysluplné technické a legislatívne obmedzenie autonómnych zbraňových systémov.

Už v júli roku 2015 bol na jednej z medzinárodných konferencií o umelej inteligencii zverejnený otvorený list, ktorý vyzýval na zákaz autonómnych zbraní. V liste sa priamo píše: „technológia umelej inteligencie dosiahla bod, v ktorom je nasadenie týchto systémov prakticky, ak nie právne, uskutočniteľné v priebehu nie desaťročí, ale rokov a v stávke je veľa: autonómne zbrane sa uvádzajú ako tretia revolúcia v zbraňových systémoch, po strelnom prachu a jadrových zbraniach“. Tento **otvorený list poukazuje na pokrok i výhody spojené s vývojom systémov AI, zvažuje riziká a dopady zneužitia LAWs na celú oblasť umelej inteligencie a graduje k výzve na „zákaz útočných autonómnych zbraní, ktoré sú mimo zmysluplnej ľudskej kontroly“**.[354]

---

dočasne odstavil mnohé ukrajinské webové stránky, nasledovala kompromitácia škodlivými kódmi HermetickWiper, IsaacWiper a HermeticWizard.
HermeticWiper: New data-wiping malware hits Ukraine. [on-line]. [cit. 10. marca 2022].
Dostupné na internete: <https://www.welivesecurity.com/2022/02/24/hermeticwiper-new-data-wiping-malware-hits-ukraine/>
IsaacWiper and HermeticWizard: New wiper and worm targeting Ukraine. [on-line]. [cit. 10. marca 2022].
Dostupné na internete: <https://www.welivesecurity.com/2022/03/01/isaacwiper-hermeticwizard-wiper-worm-targeting-ukraine/>

353 ETZIONI, ETZIONI, *Pros and Cons of Autonomous Weapons Systems* [on-line]. [cit. 5. marca 2022].
Dostupné na internete: <https://www.armyupress.army.mil/Journals/Military-Review/English-Edition-Archives/May-June-2017/Pros-and-Cons-of-Autonomous-Weapons-Systems/>

354 *Autonomous Weapons: An Open Letter from AI [Artificial Intelligence] & Robotics Researchers*. Future of Life Institute website, 28 July 2015. [on-line]. [cit. 8. marca 2022].
Dostupné na internete: <http://futureoflife.org/open-letter-autonomous-weapons/>
Okrem špičiek výskumu a vývoja umelej inteligencie mohli byť signatármi tohto listu aj ľudia z iných oblastí, a tak autor tejto publikácie mal možnosť byť tiež jedným zo signatárov.

List má pôsobivý zoznam signatárov, medzi ktorými sú okrem iných Elon Musk (vynálezca a zakladateľ spoločnosti Tesla), Steve Wozniak (spoluzakladateľ spoločnosti Apple), fyzik Stephen Hawking[355] (Univerzita v Cambridge) a Noam Chomsky (Massachusettský technologický inštitút). List podpísalo aj viac ako tritisíc výskumníkov v oblasti umelej inteligencie a robotiky...

Spolu s autormi článku *Pros and Cons of Autonomous Weapons Systems* len pripomíname to, čo sme pri opise LAWs už skôr uviedli: často nie je ľahké rozlíšiť, či ide o zbraň útočnú alebo obrannú (prieskumnú, atď.). V mnohých prípadoch aj neútočné smrtiace autonómne zbraňové systémy môžu pracovať v režime, ktorý by podľa vyššie uvedeného listu mohol byť eticky neprijateľný.

Etické výzvy idú ruka v ruke s právnymi aspektami, s čím súvisí i problém zodpovednosti pri nasadení autonómnych zbraňových systémov. Etik Robert Sparrow upozorňuje na tento eticko-právny problém tým, že základná podmienka medzinárodného humanitárneho práva alebo *jus in bello* vyžaduje, aby za úmrtia civilistov niesla zodpovednosť nejaká osoba.[356] Akákoľvek zbraň alebo iný vojnový prostriedok, ktorý znemožňuje určiť zodpovednosť za obete, túto požiadavku nespĺňa, a preto by sa nemal používať vo vojnovom konflikte.[357]

Keďže zbraňové systémy vybavené umelou inteligenciou sa rozhodujú samy, je ťažké určiť, či chybné rozhodnutie je spôsobené priamo ľudským faktorom, alebo chybami v algoritmoch AI či v autonómnom uvažovaní týchto tzv. inteligentných strojov. Predstavme si napríklad situáciu, v ktorej autonómne vozidlo porušilo rýchlostné limity tým, že sa

355 Nielen v tejto oblasti AI, ale aj vo viacerých ďalších Stephen Hawking varuje: „**It will either be the best thing that's ever happened to us, or it will be the worst thing. If we're not careful, it very well may be the last thing.**“
HIGGINS, A. *Stephen Hawking's final warning for humanity: AI is coming for us*. [on-line]. [cit. 13. marca 2022].
Dostupné na internete: <https://www.vox.com/future-perfect/2018/10/16/17978596/stephen-hawking-ai-climate-change-robots-future-universe-earth>

356 Niektorí protagonisti hypotetickej všeobecnej a uvedomelej AI smelo hovoria o nej ako o osobe. Akékoľvek vnímanie AGI ako osoby by prinieslo celé spektrum neľahkých etických a právnych výziev. Por. THURZO, V. *The Influence of Existentialism and Subjectivism on the Concept of the Human Person*. In *Spiritual and Social Experience in the Context of Modernism and Postmodernism*. Morrisville, 2021, s. 7-34.

357 SPARROW, R. *Killer Robots*. In: *Journal of Applied Philosophy.* 2007, 24, č. 1, s. 62–77.

pohybovalo príliš pomaly na diaľnici, pričom následne nebolo jasné, komu by mala byť udelená pokuta...[358] V situáciách, keď je rozhodnutie o použití smrtiacej sily na človeku, existuje jasný reťazec zodpovednosti, ktorý sa tiahne od toho, kto skutočne „stlačil spúšť", až po veliteľa, ktorý vydal rozkaz. V prípade autonómnych zbraňových systémov takáto istota a jednoznačnosť neexistuje. **Nie je jasné, kto alebo čo má niesť vinu alebo zodpovednosť**.[359]

Už v roku 2013 skupina inžinierov, odborníkov na umelú inteligenciu a robotiku a ďalších vedcov a výskumníkov z tridsiatich siedmich krajín sveta vydala „Výzvu vedcov na zákaz autonómnych smrtiacich robotov". V tejto výzve sa uvádza, že chýbajú vedecké dôkazy o tom, že by roboty mohli v budúcnosti disponovať „funkciami potrebnými na presnú identifikáciu cieľa, situačné povedomie alebo rozhodnutia týkajúce sa primeraného použitia sily". LAWs tak môžu spôsobiť vysokú mieru vedľajších škôd. Vyhlásenie sa končí zdôraznením, že „**rozhodnutia o použití hrubej sily sa nesmú delegovať na stroje**".[360]

Obava z delegovania rozhodovania o živote a smrti na neľudské subjekty sa v rámci autonómnych zbraňových systémov osobitne týka **LAWs, ktoré sú schopné vybrať si vlastné ciele**. Z tohoto dôvodu i rešpektovaný počítačový vedec Noel Sharkey vyzval na zákaz „smrtiaceho autonómneho zameriavania/cielenia", pretože porušuje zásadu rozlišovania. Zásada rozlišovania sa považuje za jedno z najdôležitejších pravidiel ozbrojených konfliktov – nie je jednoduché pre autonómne zbraňové systémy rozlíšiť, kto je civilista a kto aktívne bojuje (napr. pri mestských bojoch je to ťažké určiť aj pre ľudí).[361] **Umožnenie systémom AI rozhodovať o zameraní cieľa s najväčšou pravdepodobnosťou povedie k civilným obetiam a neprijateľným vedľajším škodám**.

V zásade ide o tie isté argumenty i závery, ktoré uvádzame naprieč celou 2. kapitolou

---

358 Por. ETZIONI, A., ETZIONI, O. *Keeping AI Legal*. In: *Vanderbilt Journal of Entertainment & Technology Law* [on-line]. 2016, 19, č. 1, s. 133–146. [on-line]. [cit. 8. marca 2017]. Dostupné na internete: <http://www.jetlaw.org/wp-content/uploads/2016/12/Etzioni_Final.pdf>

359 ETZIONI, ETZIONI, *Pros and Cons of Autonomous Weapons Systems* [on-line]. [cit. 5. marca 2022]. Dostupné na internete: <https://www.armyupress.army.mil/Journals/Military-Review/English-Edition-Archives/May-June-2017/Pros-and-Cons-of-Autonomous-Weapons-Systems/>

360 *Scientists' Call to Ban Autonomous Lethal Robots*. ICRAC, October 2013. [on-line]. [cit. 8. marca 2022]. Dostupné na internete: <http://www.icrac.net/>

361 SHARKEY, N. *Saying 'No!' to Lethal Autonomous Targeting*. In: *Journal of Military Ethics.* 2010, 9, č. 4, s. 369–383.

o limitoch a rizikách súčasných systémov AI:

- „nevyspytateľnosť“ čiernej skrinky a rizikové faktory systémov umelej inteligencie (kapitola 2.1.)
- riziká spojené s procesnými útokmi na technológie umelej inteligencie (2.2.)
- realita rôznorodosti kybernetických útokov a reálna nemožnosť vytvoriť absolútne bezpečný systém (2.3.)
- technologická komplexnosť a nutnosť fungujúcej infraštruktúry (2.4.)
- toxické psychologické a sociologické dôsledky dopadajúce od jednotlivcov až po celú spoločnosť (2.5.)
- riziká špecifického zneužitia v nasadení, ktoré dávno presahuje pôsobnosť spravodajských služieb (2.6.)
- hrozby, ktoré sú spojené so smrtiacimi automatizovanými zbraňovými systémami, ktoré vďaka algoritmom AI dokážu prekonávať schopnosti človeka, no pravdepodobne bez možnosti reálne do nich implementovať etické pravidlá a legislatívne obmedzenia (2.7.)

A prichádzame k tomu istému záveru (už skôr sme ho formulovali ako požiadavku, ktorou sme končili kapitolu 2.4.): **nutnou podmienkou prevádzky ľubovoľného systému AI, ktorý môže predstavovať riziko pre akúkoľvek ľudskú osobu, je schopnosť a možnosť človeka prebrať kedykoľvek kontrolu nad týmto systémom, resp. právo a možnosť verifikovať a prehodnotiť výsledky jeho činnosti**.

Naše závery sú kompatibilné s myšlienkovým rámcom, ktorý zastávajú Sharkey, Sparrow. signatári vyššie uvedeného otvoreného listu a ďalší. Ide o návrh na stanovenie limitov pre technologický vývoj autonómnych zbraňových systémov a vytýčenie červených čiar a mantinelov, ktoré by budúci technologický vývoj nemal prekročiť. Len pripomíname, že tieto **limity, regulácia a obmedzenia LAWs by mali predstavovať etický rámec stanovený na základe morálnych hodnôt ľudskej spoločnosti**.

Viacerí odporcovia tohoto prístupu (prevažne z armádnych a politických kruhov) vidia v tomto postoji zbytočné a neúmerné obmedzovanie vývoja a nasadenia autonómnych smrtiacich systémov. Dávajú skôr prednosť „následnej regulácii“, pri ktorej by sa mal uplatňovať vyčkávací prístup a regulácia by sa mala odvíjať od toho, ako sa objavujú nové

pokroky vo vývoji a možnostiach nasadenia LAWs. V zásade špekulujú o vývoji etiky a hodnotového rámca, prispôsobiac sa tak stavu vývoja zbraňových systémov na báze umelej inteligencie. Právni vedci, ako napríklad Kenneth Anderson a Matthew Waxman, ktorí tento prístup obhajujú, tvrdia, že regulácia bude musieť vznikať priebežne spolu s technológiou a domnievajú sa, že etika a mantinely morálnej obhájiteľnosti sa budú vyvíjať spolu s technologickým rozvojom. **Podporovatelia „následnej regulácie“ sa tak ľahko môžu dostať do oblasti hodnotového a morálneho relativizmu, snažiac sa prehodnotiť argumenty o nenahraditeľnosti ľudského svedomia a morálneho úsudku**.[362]

Zástancovia „následnej regulácie“ predpokladajú dosiahnutie tohoto cieľa prirodzeným vývojom, v rámci ktorého si ľudia budú postupne zvykať na stroje vykonávajúce funkcie s dôsledkami ohrozujúcimi život až po prípadnú smrť (napríklad riadenie automobilov alebo vykonávanie chirurgických operácií a pod.). Spoločnosť tak postupne začne akceptovať niečo podobné pri začleňovaní technológií umelej inteligencie do výzbroje.[363]

V tomto kontexte Anderson a Waxman odmietajú nami zastávaný etický rámec, regulácie a pevne stanovené podmienky pre vývoj a prevádzku LAWs. Prikláňajú sa skôr k vytvoreniu nejakého „komunitného riešenia“ s usmerneniami a zásadami, vyjadrujúcimi očakávania spoločnosti ohľadom legálneho a eticky vhodného správania a fungovania týchto systémov.[364]

I keď sme zástancami jasných etických pravidiel a noriem vyplývajúcich z morálnych hodnôt, niektoré postrehy podporovateľov „**následnej regulácie**“ nám môžu byť nápomocné pri uchopení etických výziev a problémov, ktoré sú a budú súčasťou extrémne rýchleho vývoja systémov umelej inteligencie, a to nielen v oblasti vojenského využitia.

Na margo eticko-právneho diskurzu, z ktorého veľká časť prebieha v armádnych kruhoch USA, treba povedať, že súčasné nastavenie je priaznivé realizácii jasného etického rámca a pravidiel. Vládny Defense Innovation Board[365] v roku 2019 v dokumente *AI Principles:*

362 Por. ANDERSON, K., WAXMAN, M. C. *Law and Ethics for Autonomous Weapon Systems: Why a Ban Won’t Work and How the Laws of War Can*. Stanford University, Hoover Institution Press, Jean Perkins Task Force on National Security and Law Essay Series, 2013.

363 Podobnosť s inými oblasťami erózie hodnôt je úplne náhodná...

364 Por. ANDERSON, K., WAXMAN, M. C. *Law and Ethics for Robot Soldiers*. In: *Policy Review.* 2012, 176, s. 46.

365 Dostupné na internete: <https://innovation.defense.gov/>

Recommendations on the Ethical Use of Artificial Intelligence by the Department of Defense[366] navrhol princípy využívania systémov umelej inteligencie v americkej armáde, pričom je pre tieto princípy podstatné, že rozhodovacia právomoc zostáva na človeku, osobitne ak ide i misie s použitím smrtiacej sily.[367]

Medzi podstatné návrhy uvedené v dokumente sa radí:

- už uvedená **zodpovednosť** (rozhodovacia právomoc) pri nasadení, v rámci ktorej by vojenský personál mal uplatňovať primeranú úroveň úsudku a zostať zodpovedný za vývoj, nasadenie, používanie a dôsledky využívania systémov AI.
- požiadavka na extrémnu **opatrnosť pri príprave trénovacích dát**, aby sa systémy AI vyhli neúmyselnej zaujatosti/predsudkom (unintended bias), ktorých dôsledkom by boli nesprávne rozhodnutia v operačnom nasadení.[368]
- **dosledovateľnosť**, v rámci ktorej musí existovať možnosť vrátiť sa späť k akémukoľvek dielčiemu výstupu alebo procesu a umožniť tak analýzu rozhodovacieho procesu algoritmov AI. K dosledovateľnosti prináleží požiadavka na kontrolovateľné metodiky, zdroje údajov, postupy a dokumentácie dizajnov systémov.[369]
- **spoľahlivosť** ako výsledok presne definovanej oblasti využitia, zodpovedajúcej dizajnu a exaktnému vývoju systému AI, v kombinácii s celou paletou prísnych testov verifikujúcich spoľahlivosť v tejto definovanej oblasti.
- **ovládateľnosť**, t.j. schopnosť odhaľovať a brániť neúmyselnému poškodeniu alebo narušeniu činnosti systému AI kombinovaná s reálnou možnosťou zastavenia či prerušenia činnosti v prípade detekovaných problémov.

366 *AI Principles: Recommendations on the Ethical Use of Artificial Intelligence by the Department of Defense*. [on-line]. [cit. 9. marca 2022].
Dostupné na internete: <https://admin.govexec.com/media/dib_ai_principles_-_supporting_document_-_embargoed_copy_(oct_2019).pdf>

367 TUCKER, P. The Pentagon's AI Ethics Draft Is Actually Pretty Good. [on-line]. [cit. 9. marca 2022].
Dostupné na internete: <https://www.defenseone.com/technology/2019/10/pentagons-ai-ethics-draft-actually-pretty-good/161005/>

368 Tzv. predsudky systémov AI ako dôsledok nesprávne zvolenej, či nekvalitnej množiny trénovacích dát sme rozoberali v kapitole 2.1.2.

369 Ide o našu známu transparentnosť, aj keď treba povedať, že vzhľadom na fenomén black boxu tažko uveriť, že pri skutočne zložitých systémoch AI by toho vojenskí špecialisti boli schopní.

I keď sme to priamo zatiaľ neuviedli, zo všetkých oficiálnych aktivít na poli nasadenia vojenských systémov umelej inteligencie vyplýva, že **jednotlivé krajiny sa nedokážu vzdať tak lákavej technologickej výhody a sú pevne rozhodnuté technológie umelej inteligencie implementovať v celej šírke možného zmysluplného využitia**.

Navrhnuté pravidlá pre využívanie vojenských systémov umelej inteligencie v armáde USA, ktorá je prakticky najďalej s vývojom a nasadením technológií AI, poukazujú tiež na **jasný zámer definovať a implementovať pevný etický rámec**, bez ktorého by sa rozmachom armádnych systémov AI otvorila Pandorina skrinka.

Vzhľadom na spektrum štátov, ktoré pracujú na aktívnom nasadení technológií umelej inteligencie vo vojenskej oblasti, **je však diskutabilné, či analogický záujem o striktný etický rámec bude existovať aj u ostatných**.

Na základe doterajších skúseností v oblasti limitov a rizík súčasných systémov AI **je tiež otázne, či v celom spektre technológií a algoritmov umelej inteligencie bude možné pevný etický rámec dodržať**. Vieme si to predstaviť pri cielenom – čo môže znamenať aj limitujúcom a tým aj znevýhodňujúcom – dizajne systémov AI s dôrazom na dodržiavanie navrhnutého rámca (ethics by design), avšak pri súčasnej snahe o čo najširšie adoptovanie umelej inteligencie v armáde a získanie konkurenčnej výhody sme v tomto smere skeptickí.

Ako dovetok tejto kapitoly môžeme uviesť vyjadrenie ruského ministra obrany Sergeja Šojgu v jednom z interview na medzinárodnom vojenskom a technologickom fóre Army-2020: „Už to nie sú veci, ktoré vyzerajú ako hračky z detského sveta, je to niečo, čo predstavuje vážne nebezpečenstvo a seriózny zbraňový systém, pokiaľ je to reálne riadené neurónovými sieťami a umelou inteligenciou...“[370]

Stať o systémoch umelej inteligencie v armáde je písaná v prvých týždňoch vojnového konfliktu na Ukrajine. Zo správ, ktoré prichádzajú z brífingov velení priamo či nepriamo zúčastnených strán tohoto konfliktu i z monitoringu na rôznych platformách zverejňovanej

370 От робототехники до беспилотников: Шойгу рассказал о новинках ОПК на форуме «Армия-2020». [on-line]. [cit. 1. marca 2022].
Dostupné na internete: <https://youtu.be/U8V4OZyNSac?t=152>
Príkladom takýchto systémov, ktoré Rusko vo vojnovom konflikte na Ukrajine nasadilo, sú i umelou inteligenciou vybavené drony KUB a Lancet od Zala Aero Group, ktorá je súčasťou koncernu Kalašnikov.

komunikácie členov armád a bojujúcich skupín[371] badať veľmi nešťastnú kalkuláciu, ktorá tento v určitých aspektoch asymetrický konflikt, doplnený okrem bežnej vojenskej techniky aj niektorými skutočne modernými zbraňovými systémami, sprevádza: civilné obyvateľstvo a bezbranní ľudia sa používajú ako súčasť obrannej i útočnej taktiky a sú vystavení realite, v ktorej na prvom mieste je čokoľvek iné, než dobro človeka.

Tento fenomén je známy i z rôznych konfliktov na blízkom východe. Ak ho naviac skombinujeme s nasadením smrtiacich automatizovaných zbraňových systémov, ohrozenie ľudských bytostí môže smerovať ku katastrofe.

2.8. Môžeme umelej inteligencii dôverovať?

Podľa štúdie zverejnenej tímom odborníkov na stránkach Scientific Reports algoritmy strojového učenia začali zasahovať do úloh, ktoré boli tradične vyhradené pre ľudský úsudok, a sú čoraz schopnejšie podávať dobré výkony v nových a náročných úlohách.

Na troch rôznych experimentoch tento tím ukázal, že ľudia sa s rastúcou náročnosťou úloh viac spoliehajú na algoritmické rady než na interakciu okolia. Ľudské subjekty mali tiež tendenciu výraznejšie ignorovať nepresné rady označené ako algoritmické v porovnaní s rovnako nepresnými radami označenými ako rady pochádzajúce od skupiny kolegov. Výrazne sa prejavil i vplyv sociálnych médií, online recenzií alebo osobných sociálnych sietí – tento vplyv je jednou z najdôležitejších síl, ktoré ovplyvňujú rozhodovanie jednotlivcov.[372]

Javí sa, že pokiaľ prejavy a výstupy systémov umelej inteligencie spadajú do rámca exaktnosti, logiky a matematickej presnosti, t.j. do intelektuálneho rámca, ktorý sme si podvedome vytvorili, resp. prijali pre vedu a techniku, sme náchylní týmto systémom a priori dôverovať.[373]

Nerozporujúc úžasný potenciál a pokrok vo vývoji i úspechy v nasadení systémov ANI sme

371 Sociálna sieť Telegram si v tomto priamo žiada analytické nasadenie systému AI – medzi propagandou a dezinformáciami najrozličnejšieho druhu sa totiž nájde aj veľa faktických, operačných i taktických informácií z frontovej línie.

372 BOGERT, E., SCHECTER, A., WATSON, R. T. *Humans rely more on algorithms than social influence as a task becomes more difficult*. In: *Sci Rep.* [on-line]. 2021, 11, 8028. [cit. 20. februára 2022]. Dostupné na internete: <https://doi.org/10.1038/s41598-021-87480-9>

373 Naviac, táto dôvera rastie úmerne náročnosti úloh, ktoré s pomocou týchto systémov máme realizovať, a to v kontexte postmodernej rezignácie na rácio neveští nič dobrého.

sa však v rámci tejto kapitoly potýkali s rôznymi chybami, limitmi a rizikami jej rôznych druhov a viacerých scenárov nasadenia.

Preto v kontexte rastúcej dôvery v systémy umelej inteligencie a vnímajúc problémy, s ktorými sa nasadenie týchto systémov potýka i oblastí, ktorých sa ich využitie dotýka, môžeme a musíme stanoviť podmienky, bez splnenia ktorých by nasadenie systémov AI do reálneho sveta, v ktorom interagujú s človekom a vplývajú na spoločnosť, nemalo byť umožnené.

Systémy umelej inteligencie vo všeobecnosti musia byť:

- **legálne** – vyhovovať požadovaným normám, zákonom a reguláciám
- **etické** – spĺňať požadované etické kritériá
- **bezpečné** – dosahovať potrebné štandardy bezpečnosti a robustnosti

Len takto sa môžeme priblížiť k cieľu, ktorým je **dôveryhodná umelá inteligencia**. Hovoríme tak o **systémoch AI, ktoré sú zamerané na človeka** a podporujú zodpovedné riešenia dbajúce na ľudské potreby, bezpečnosť a súkromie.[374]

374 *Trustworthy AI is human-centered*. [on-line]. [cit. 20. februára 2022]. Dostupné na internete: <https://www.ibm.com/watson/trustworthy-ai>

3. Umelá inteligencia v optike etiky

Existujú len dva druhy priemyslu, ktoré svojim zákazníkom hovoria používatelia.
Nelegálne drogy a softvér.[375]

Pohybujúc sa už viac než dve desaťročia v oblasti kybernetickej bezpečnosti si uvedomujeme, že prakticky ešte nikdy nebolo badať toľké obavy z nasadenia novej technológie a tak seriózny záujem[376] o etické otázky ako v prípade systémov umelej inteligencie. Kľúčové slovo či značka (hashtag) #AIEthics v komunite odborníkov neoznačuje okrajovú záležitosť, ktorá je mimo zorného uhľa pohľadu tých, čo umelej inteligencii skutočne rozumejú, ale stáva sa súčasťou hlavného prúdu vývoja, implementácie a používania systémov AI v rámci reálneho sveta.

Aby sme si rozumeli, s futurologickými obavami takmer na úrovni sci-fi sa spoločnosť stretáva už od pionierskych čias formovania konceptu umelej inteligencie a prvých krokov na poli vývoja týchto systémov. V našom (a nielen našom) prípade však ide o seriózny záujem celej spoločnosti, ktorá je konfrontovaná s technológiami, ktoré majú potenciál hlboko zasahovať a ovplyvňovať jednotlivcov i celé spoločenské celky.

Tento seriózny záujem sa neobmedzuje len na oblasť umelej inteligencie, ale má oveľa širší záber, ktorý koreluje a má i kauzálnu súvislosť s treťou priemyselnou revolúciou, digitálnym vekom a zmenou paradigmy nastupujúcej informačnej spoločnosti, v rámci ktorej sa na základe zmien v technológiách a vzhľadom na prevratný vedecký i technologický pokrok mení medziľudská komunikácia vo svojej podstate, menia sa vzťahy, mení sa spoločnosť a princípy jej fungovania.[377]

Prvý zviditeľnením tohoto záujmu bol na verejnosť „prevalený“ zápas o ochranu osobných

375 Prof. Edward R. Tufte.

376 Už skôr boli komunikované témy a vybojované zápasy, bez ktorých by sme si moderný svet informačných technológií nevedeli predstaviť, napr. zápas open-source vs. closed software, legitimita slobodných licencií typu Creative Common a pod. Vždy však išlo z pohľadu spoločnosti len o okrajovú záležitosť.

377 Por. ŠANTAVÝ, *Analýza virtuálneho sveta v kontexte náboženskej formácie, pôsobenia a života Cirkvi,* [on-line], s. 20. [cit. 20. februára 2022].
Dostupné na internete:
<https://peter.santavy.cloud/data/uploads/docs/analyza_virtualneho_sveta_v_kontexte_nabozenskej_formacie_posobenia_a_zivota_cirkvi.pdf>

údajov a jasné pravidlá pre sledovanie a monitoring občanov i boj o prístup k informáciám s cieľom angažovať sa v oblasti ľudských práv.[378] Išlo a ide o zápas, ktorý prináša svoje ovocie, či už ide o rast povedomia a verejnú diskusiu o ochrane osobných údajov a práve na súkromie, alebo o legislatívne rámce a regulácie, z ktorých najznámejšou a pravdepodobne i najúčinnejšou je Nariadenie Európskeho parlamentu a Rady (EÚ) 2016/679 (General Data Protection Regulation – GDPR) o ochrane fyzických osôb pri spracúvaní osobných údajov a o voľnom pohybe takýchto údajov.[379]

Ďalším krokom na ceste upevnenia záujmu celej spoločnosti o problémy, ktoré ju bytostne ovplyvňujú, je akcentovanie etického rozmeru širokej oblasti umelej inteligencie. Skúsme sa tomu venovať aj v tejto kapitole.

3.1. Etické výzvy prameniace z limitov a rizík umelej inteligencie

Rozoberajúc limity a riziká súčasných systémov umelej inteligencie v druhej kapitole, sme nie raz vyvodzovali závery, ktoré možno naplno zaradiť do agendy etiky AI.

Pripomeňme si aspoň tie podstatné:

- [380]**Súčasné systémy AI sprevádzajú oprávnené a vážne obavy z toho, že ak nechápeme ako tieto systémy pracujú, nemôžeme im reálne dôverovať a ťažko dokážeme predpovedať okolnosti, za ktorých tieto systémy zlyhajú.**

  Keďže v zásade nevieme, čo presne sa neurónová sieť naučila a ako spoľahlivo to dokáže aplikovať, prakticky každý sofistikovaný systém umelej inteligencie sa javí ako *black box* – čierna skrinka, ktorá niečo vykonáva, ale ako a prečo tak robí, nemusí byť jasné.

  Naviac celkový dizajn, voľba algoritmov a nastavenie hyperparametrov, prípadne

378 Náhľad do týchto problémov prináša napríklad analýza viacerých káuz z oblasti úniku a zverejňovania tajných informácií (whistleblowing) – Cablegate z roku 2010 a Snowden vs. NSA začínajúca sa v roku 2013 s dosahom na celé posledné desaťročie, prípadne viaceré škandály Facebooku ohľadom zneužitia a monetizácie osobných údajov používateľov.
Por. ŠANTAVÝ, *Niektoré výzvy informačnej spoločnosti v oblasti morálnej teológie*, s. 117-135.

379 *Nariadenie Európskeho parlamentu a Rady (EÚ) 2016/679 z 27. apríla 2016 o ochrane fyzických osôb pri spracúvaní osobných údajov a o voľnom pohybe takýchto údajov, ktorým sa zrušuje smernica 95/46/ES (všeobecné nariadenie o ochrane údajov)*. [on-line]. [cit. 19. februára 2022]. Dostupné na internete: <https://eur-lex.europa.eu/legal-content/SK/TXT/?uri=celex%3A32016R0679>

380 Kapitola 2.1.1.

niektoré ďalšie aspekty návrhu tých najlepších systémov AI sú postavené viac na odbornej intuícii a metóde pokus-omyl, než na exaktných postupoch, ktoré by mohli byť univerzálne aplikovateľné a tým ja vo väčšej miere verifikovateľné.

Problém s verifikovateľnosťou postupov a výsledkov činnosti systémov umelej inteligencie sa tak premieta i do oblasti implementácie etických pravidiel, či už ide o spôsob, ako hodnotiť a kontrolovať činnosť systémov AI, alebo ide priamo o spôsob implementácie regulácií technológiami AI.[381]

- [382]Súčasné systémy umelej inteligencie trpia rôznymi zraniteľnosťami. **Musíme mať neustále na pamäti, že tieto systémy môžu zlyhávať najrozličnejšími a často neočakávanými spôsobmi** v dôsledku nemožnosti pripraviť dostatočne veľkú množinu trénovacích dát, prípadne ich nesprávnej voľby bez dostatočnej kvality alebo s predsudkami, nadmernému prispôsobovaniu sa tréningovým údajom, efektu dlhého chvosta, rizikám plynúcim z oklamania hlbokých sietí, ich zraniteľností a tzv. povier, pod čo sa môže podpísať i nedostatok odbornej erudovanosti potrebnej pre dizajn a prípravu funkčného a úspešného riešenia i ladenia hyperparametrov.

  Nielen riziká priameho zlyhania, ale aj realita výsledkov, ktoré môže byť ťažké správne interpretovať (čo sa vlastne sieť naučila, čo výstup z daných dát na vstupe v skutočnosti znamená) a neschopnosť predvídať, kedy sa jednotlivé zlyhania prejavia (za akých podmienok, pri akej súhre okolností, s dôsledku akej dynamiky vnútorného vývoja, resp. činnosti systému AI).

  S týmito zlyhaniami a obmedzeniami treba rátať aj v oblasti snahy o technologickú implementáciu etických noriem a mantinelov. Nemusí to byť vôbec triviálne a v niektorých scenároch nasadenia ani uskutočniteľné.

- **[383]Systémy umelej inteligencie robia chyby, ktoré sú diametrálne odlišné od ľudských chýb a zlyhaní**. Preto môžu byť prekvapivé, neočakávané a nebezpečné.

  Naviac, **ak by sme priamo algoritmami AI implementovali regulačné**

381 Vo viacerých oblastiach systémov AI ani nie je možné technologicky tieto dva rozmery implementácie od seba oddeliť – v praxi môže ísť o súčasť funkcionality, algoritmického návrhu a procesu učenia sa celého systému.

382 Kapitola 2.1.2.

383 Kapitola 2.1.2.

mechanizmy, musíme rátať s rovnakým rizikom nečakaných pochybení týchto mechanizmov počas prevádzky.

- [384]Vzhľadom na rozličné typy útokov zameraných na procesy umelej inteligencie sa **etické výzvy neviažu len na návrh a realizáciu systémov AI, no rozširujú sa aj o oblasť etiky použitia a z toho prameniace riziká zneužitia**.

  Ide teda o pohľad na **etický rozmer zodpovedného vývoja systémov a procesu učenia**, pri ktorom je jednou zo základných požiadaviek princíp „**etika už od návrhu“ (ethics by design)**.

  Minimálne v rovnakej miere ide aj o pohľad na **etický spôsob využívania technológii AI**, t.j. bez akejkoľvek snahy o ich kompromitáciu, resp. zneužitie proti iným, čo sa v nemalej miere premieta aj do problematiky kapitol 2.5. až 2.7.

- [385]**Systémy umelej inteligencie sú vystavené aj problémom v oblasti kybernetickej bezpečnosti**, keďže mnohé riziká atakujúce bezpečnosť procesov umelej inteligencie, zneužívajúce nedostatky dizajnu alebo amplifikujúce dôsledky rizikových faktorov, ako vektor útoku používajú zraniteľnosti v oblasti kybernetickej bezpečnosti. Kybernetická bezpečnosť je proces, ktorý má svoju dynamiku a vyjadruje skutočnosť, že **ani v oblasti systémov umelej inteligencie neexistuje dokonale bezpečný a spoľahlivý systém**.

  S rozvojom zložitých systémov AI treba mať neustále na pamäti, že počet zraniteľností je priamo úmerný komplexnosti systémov. Podobne i množstvo kompromitácií, resp. zneužití týchto zraniteľností priamo rastie s mierou nasadenia a využívania v reálnom svete.

  Okrem opakujúcej sa požiadavky na etický rozmer zodpovedného vývoja a etického využívania systémov AI treba pamätať aj na relatívnu bezpečnosť a spoľahlivosť týchto systémov. Akákoľvek ich absolutizácia môže viesť nielen k ekonomickým škodám, ale v niektorých oblastiach aj k ohrozeniu človeka a spoločnosti, ktoré môže byť – v pohľade na možné spektrum nasadenia od autonómnych vozidiel až po algokraciu – fyzické, psychologické, morálne, právne,...

- [386]Krátky náčrt problematiky technologického vybavenia a nutnej infraštruktúry

384 Kapitola 2.2.

385 Kapitola 2.3.

pre nasadenie systémov AI vo viacerých oblastiach reálneho sveta, ktorý sme uviedli v kapitole 2.4., poukazuje na veľkú komplexnosť týchto systémov. A **komplexnosť je problém – je nielen rizikovým faktorom bezpečnosti a stability fungovania systémov, ale mnohokrát i neľahkou výzvou pre úspešnú a jasnú implementáciu regulácií a etických pravidiel**.

- [387]Úspešné nasadenie mnohých moderných systémov AI vyžaduje prísun extrémneho množstva dát z reálneho sveta a priamo z ľudského prostredia. Ľudia i celá spoločnosť sú preto vystavení neustálemu dohľadu a sledovaniu, ktoré je však len veľmi málo pod kontrolou, ak vôbec. **Pri veľkých systémoch zhromažďujúcich denno denne extrémne množstvo dát je veľmi ťažké zaručiť etiku spracúvania týchto údajov a vylúčiť riziko ich zneužitia**.

  Systémy umelej inteligencie v oblasti médií a sociálnych sietí posúvajú paradigmatickú zmenu informačnej spoločnosti na vyššiu úroveň: miesto informácií sa základnou komoditou stávajú priamo ľudia, keďže prichádza k ovplyvňovaniu ľudského vnímania, rozhodnutí i konania a k neustálemu zlepšovaniu schopností systémov AI vytvárať modely predikujúce konanie konkrétneho človeka. **Systémy umelej inteligencie sa s veľkou mierou istoty stávajú schopnými vytvárať veľmi presné psychologické profily, odhaľovať akékoľvek väzby a mnohé osobné informácie, predpovedať a ovplyvňovať naše konanie**. **Tieto systémy však mnohokrát nie sú predmetom regulácie, resp. verejnej kontroly.**

  Ak dosiahnutý efekt a vplyv môže byť tak závažný a prakticky takmer istý, existuje riziko i pokušenie využiť to, presnejšie – manipulovať a zneužiť (či už vo svete biznisu, ale aj v oblastiach oveľa kritickejších).

  Neuvedomujúc si, ako je naša myseľ a psychika zraniteľná, tak **postupne prechádzame od technologického prostredia založeného na systémoch AI k prostrediu založenému na závislosti a manipulácii**.

  Systémy AI sa podieľajú na stieraní hranice medzi pravdou a lžou, medzi realitou a fiktívnym svetom. Prichádza tak k extrémnemu rozdeleniu spoločnosti a narušeniu vzťahov, k erózii hodnôt a ideologickej divergencii, k nárastu zatvrdilej nevedomosti, ignorancii faktov a relativizácii pravdy, k neschopnosti riešiť aktuálne

386 Kapitola 2.4.

387 Kapitola 2.5.

civilizačné výzvy, ku kríze demokracií a spoločenských mechanizmov, k ekonomickým dôsledkom a pre mnohých i k potencionálnemu smerovaniu k civilizačnému kolapsu...

Míľnikom, ktorého by sme sa mali obávať, teda nie je budúca technologická singularita v oblasti umelej inteligencie, v ktorej AI prevýši náš intelekt, ale oveľa skôr moment, keď technológia ovládne a prekoná naše slabosti... už vtedy prichádza víťazstvo umelej inteligencie a porážka ľudstva.

- [388]Technológie samy o sebe nie sú tou priamou hrozbou - dokážu však prebudiť v človeku aj v spoločnosti tie najhoršie stránky ľudského bytia, ktoré sa tak stávajú existenčnou hrozbou.

  V kontexte slabej umelej inteligencie (ANI) ešte stále konečnú voľbu cieľov nerobia stroje, ale človek, takže stále je v ľudskej moci tieto dôsledky ovplyvniť a zmeniť. Preto musia existovať etické pravidlá a regulácie, ktoré by dokázali chrániť jednotlivca i celú spoločnosť voči rizikám sofistikovanej práce systémov AI v rámci sociálnych sietí, systémov riadenia spoločnosti, resp. analogických platforiem.

- [389]**Viaceré aplikácie systémov umelej inteligencie v reálnom svete budú musieť v zlomkoch sekúnd riešiť rôzne kritické situácie – a vedieť ich vyriešiť eticky**. Napríklad rozhodovanie autopilota plne autonómneho vozidla medzi potencionálnym ohrozením života posádky alebo ostatných účastníkov v čase blížiacej sa alebo prebiehajúcej dopravnej nehody.

  Mnohé z týchto systémov sú a budú využívané v oblastiach, kde môže byť ohrozené zdravie a život človeka. Základné etické normy pre ich prevádzku musia byť legislatívne ustavené. Otázkou je, **do akej miery tieto eticko-právne regulácie budú aj technicky úspešne implementované**.

- [390]Vo viacerých oblastiach nasadenia (napr. autonómne vozidlá, LAWs) bude **treba vyriešiť rozdelenie zodpovednosti medzi človekom a riadiacim systémom AI** i s tým súvisiace legislatívne požiadavky pre reálnu prevádzku týchto systémov.[391]

388 Kapitola 2.5.

389 Kapitola 2.5.

390 Kapitola 2.5.

391 Napríklad v oblasti nasadenia autonómnych vozidiel do bežnej prevádzky sa diskutujú legislatívne

- [392]Ako stanoviť **akceptovanú mieru neistoty v oblasti spoľahlivosti a robustnosti** systémov umelej inteligencie?
- [393]**Fenomén digitálneho rozdelenia (digital divide)**, ktorý sa s neustálym rozvojom informačných technológií a prechodom k informačnej spoločnosti stáva reálnym problémom, môže byť vďaka necitlivému nasadeniu technológií umelej inteligencie oproti súčasnosti ešte znásobený. Či už ide o dopady v oblasti života spoločnosti (2.5.) alebo – a to je vážnejšie – v oblastiach súvisiacich s algoritmickým riadením spoločnosti (2.6.), napr. v dostupnosti elektronických služieb štátu, zdravotnej starostlivosti, vzdelávaní, finančných službách a pod.
- [394]**Existuje reálne riziko zneužitia systémov AI v rámci algokracie na efektívnu kategorizáciu a obmedzovanie ľudských práv občanov**. Napr. čínsky systém sociálneho kreditu.

  Technológie umelej inteligencie umožňujú aj za sprísnených legislatívnych obmedzení reálne zneužitie tajnými službami, vládnymi agentúrami a korporáciami v dohľadových, analytických a predikčných systémoch. **Aké sú možnosti nastavenia etických mantinelov a reálne dodržiavaných právnych rámcov v oblasti riadenia štátu, spravodajstva a dohľadu?**
- [395]**Trio technológií umelej inteligencie – sledovanie a dohľadové systémy spojené s analytickými nástrojmi a schopnosťou vytvárať modely predikujúce ľudské konanie či vývoj situácie – tvorí v súčasnosti skutočne silnú technologickú výbavu (nielen) spravodajských služieb**.

  Systémy umelej inteligencie, ktorým sa v súčasnosti analýza metadát, predikcia a tvorba záverov zveruje, tak de facto majú rastúci potenciál rozhodovať o bytí a nebytí, „generovať“ dôkazy, ktoré môžu byť použité napríklad pri obvinení z vlastizrady, identifikácii podozrivých osôb, dokazovaní nelegálnej činnosti a pod.

---

požiadavky na tzv. bezpečnostného operátora, hľadajúc objektívnu mieru zodpovednosti ľudského faktora pri riadení autonómneho vozidla v rámci jednotlivých stupňov automatizácie, ktoré sme uviedli v kapitole 2.5.

392 Kapitola 2.5.

393 Kapitola 2.5. a 2.6.

394 Kapitola 2.6.

395 Kapitola 2.6. a 2.7.1.

Ako realizovať striktnú zákonnosť i dôsledný dohľad demokraticky zvolených zástupcov a spoločnosti pri nasadení systémov AI v rámci spravodajských služieb, dohľadových systémov a všetkých foriem i stupňov algoritmického riadenia spoločnosti?

Uvedomujeme si, že technológie využívajúce algoritmy umelej inteligencie dokážu byť veľmi účinným nástrojom pre spravodajské služby a dohľad s takým presahom do ostatných oblastí verejného života a štátu, ktorý môže znamenať veľký posun v procesoch a spôsobe fungovania spoločnosti. Avšak vzhľadom na politický, ideologický i spoločenský kontext v rôznych častiach sveta a riziká systémov AI treba zabezpečiť, aby prostriedky umelej inteligencie neboli zneužité alebo sa vymkli kontrole.

Preto je treba v celom spektre nasadenia od spravodajských služieb až po algoritmické riadenie akcentovať veľkú rozvážnosť a vyžadovať striktnú zákonnosť i dôsledný dohľad demokraticky zvolených zástupcov, obmedziť dopady na sociálnu spravodlivosť a zabezpečiť dodržiavanie ľudských práv a hodnôt.

- [396]Pokročilé modelovanie technológií, simulácia priebehu vojenských operácií i priebehu častí konfliktu a silný virtualizačný efekt môžu viesť k morálnemu znecitliveniu obsluhujúceho personálu a velenia. Ovocím môže byť modelovanie nových druhov zbraní hromadného ničenia, návrh vojenských operácií bez vnímania strát na životoch a utrpenia civilného obyvateľstva a pod.

  Rizikovým faktorom je teda psychologický aspekt vytvorenia si podvedomého chápania boja ako virtuálnej reality. **Dôsledkom môže byť strata citlivosti pre ohrozenie ľudských životov a reálnych hodnôt pri nasadení v skutočnom boji**. Tento rizikový faktor je o to záludnejší, že sa „pod kožu“ dostáva postupne a podvedome, mnohokrát bez varovnej reakcie svedomia.

- [397]Vďaka technológiám umelej inteligencie sú vo viacerých scenároch nasadenia **autonómne zbraňové systémy lepšie ako ľudia**, pričom treba rátať s multiplikačným efektom účinnosti ich nasadenia. Preto bude obmedzovanie nasadenia týchto systémov veľmi ťažké.

396 Kapitola 2.7.2. - 2.7.4.

397 Kapitola 2.7.5.

LAWs (autonómne zbraňové systémy so smrtiacim účinkom) využívajúce algoritmy AI dokážu byť neuveriteľne technologicky adaptabilné, relatívne lacné a minimálne rovnako nebezpečné ako najmodernejšie zbraňové systémy riadené ľudskou obsluhou. **Ako dokážeme zabezpečiť, že smrtiace autonómne systémy, ktoré sú zámerne dizajnované tak, aby dokázali efektívne pracovať aj pri strate spojenia, budú vedieť správne použiť svoju smrtiacu silu a rešpektovať požadované mantinely, ak je nám známe, že systémy AI robia mnohokrát iné chyby ako ľudia a nevieme ich vytrénovať tak, aby vedeli adekvátne odpovedať na všetky reálne situácie?**

- [398]Pri skupinovom riadení autonómnych zbraňových systémov sa pýtame – je koordinované riadenie na základe interakcie jednotlivých systémov AI natoľko robustné, aby prepojené LAWs dokázali správne pracovať aj pri výpadku kritického množstva zapojených automatických systémov?

  Vieme dostatočne odhadnúť správanie centrálne riadeného (a zároveň parciálne autonómneho) alebo koordinovane riadeného skupinového automatického zbraňového komplexu vo vyhranených situáciách?

  Nehrozí, že vzhľadom na znásobenú a komplexnejšiu účinnosť celej skupiny LAWs riadených umelou inteligenciou pri zlyhaní tohto riadiaceho systému AI budú jeho jednotlivé časti konať bez obmedzení a životy ohrozujúcim spôsobom?

- [399]Vďaka nebývalému rozmachu informačných technológií, ich prieniku prakticky do všetkých oblastí života rodiacej sa informačnej spoločnosti a strategickej dôležitosti pre kritickú infraštruktúru štátov i vďaka pokročilým technológiám na poli kybernetickej bezpečnosti a kriminality, ktoré úspešne implementujú algoritmy umelej inteligencie, sa vedenie vojenských operácií v kybernetickom priestore stalo neoddeliteľnou súčasťou portfólia moderných armád.

  Technológie umelej inteligencie nasadené v obranných kybernetických systémoch sú účinnou – a treba zdôrazniť aj legitímnou – pomocou pri ochrane nielen vojenských, ale akýchkoľvek dôležitých aktív spoločnosti.

398 Kapitola 2.7.6.

399 Kapitola 2.7.7.

Avšak útočné kybernetické zbraňové systémy, ktoré svojimi doterajšími úspechmi v reálnom nasadení dokázali svoju strategickú opodstatnenosť, obnášajú riziká, ktoré treba eliminovať. Už Stuxnet ukázal, že ani tak sofistikovaná zbraň nie je úplne pod kontrolou, keďže sa rozšíril aj do iných systémov, než na ktoré bol zameraný a zaútočil aj na iné služby a organizácie. Kybernetické zbraňové systémy využívajúce algoritmy AI principiálne nie sú imúnne voči žiadnemu z rizík, ktoré sme uviedli v kapitolách 2.1. až 2.4., pritom však ich dôsledky sa prejavujú v reálnom svete – od ohrozenia konkrétnych životov cez zlyhanie dôležitej infraštruktúry až po havárie kritických systémov spoločnosti.

Preto útočné kybernetické systémy využívajúce súčasné technológie umelej inteligencie považujeme za veľmi rizikové a nevhodné (nielen) pre vojenské nasadenie.

- [400]Vzhľadom na riziká a dopady zneužitia LAWs na celú oblasť vývoja a využitia systémov umelej inteligencie sa v poslednom desaťročí stupňuje apel mnohých odborníkov z tejto oblasti **proti akémukoľvek nasadeniu útočných autonómnych zbraní, ak sú mimo zmysluplnej ľudskej kontroly**.

  V súčasnosti **nemáme vedecké dôkazy o schopnosti automatizovaných systémov disponovať funkciami potrebnými na presnú identifikáciu cieľa, situačné povedomie alebo rozhodnutia týkajúce sa primeraného použitia sily**. Umožnenie systémom AI rozhodovať o zameraní cieľa s najväčšou pravdepodobnosťou povedie k civilným obetiam a neprijateľným vedľajším škodám. **LAWs tak môžu spôsobiť vysokú mieru vedľajších škôd a preto sa rozhodnutia o použití hrubej sily nesmú delegovať na stroje**.

  Obava z delegovania rozhodovania o živote a smrti na neľudské subjekty sa v rámci autonómnych zbraňových systémov osobitne týka LAWs, ktoré sú schopné vybrať si vlastné ciele. Schopnosť strojového smrtiaceho autonómneho zameriavania a cielenia porušuje zásadu rozlišovania, ktorá sa považuje za jedno z najdôležitejších pravidiel ozbrojených konfliktov – napríklad nie je jednoduché pre autonómne zbraňové systémy rozlíšiť, kto je civilista a kto aktívne bojuje.

- [401]**Schopnosti umelou inteligenciou poháňaných autonómnych zbraňových**

400 Kapitola 2.7.8.

401 Kapitola 2.7.8.

systémov a kybernetických zbraní i napriek všetkým rizikám vedú k zvyšujúcim sa tlakom na financovanie a zavádzanie útočných kybernetických zbraní. Jednotlivé krajiny sa nedokážu vzdať tak lákavej technologickej výhody a sú pevne rozhodnuté technológie umelej inteligencie implementovať v celej šírke možného zmysluplného využitia.

Problematika obmedzenia týchto útočných systémov je skomplikovaná aj reálnym stieraním hraníc medzi obranným a útočným nasadením takmer vo všetkých oblastiach vojenského využitia technológií umelej inteligencie.

- [402]**Nutnou podmienkou prevádzky ľubovoľného systému AI, ktorý môže predstavovať riziko pre akúkoľvek ľudskú osobu, je schopnosť a možnosť človeka prebrať kedykoľvek kontrolu nad týmto systémom, resp. právo a možnosť verifikovať a prehodnotiť výsledky jeho činnosti**.

  Limity, regulácia a obmedzenia LAWs by mali predstavovať etický rámec stanovený na základe morálnych hodnôt ľudskej spoločnosti, nie na základe relativistickej tzv. „následnej regulácie“.

  Na základe uvedomenia si problémov spojených so smrtiacim využitím algoritmov AI a angažovanosti odbornej verejnosti si vlády a armády viacerých krajín uvedomujú potrebu regulačného rámca. Napríklad navrhnuté pravidlá pre využívanie vojenských systémov umelej inteligencie v armáde USA, ktorá je prakticky najďalej s vývojom a nasadením technológií umelej inteligencie, poukazujú i na jasný zámer definovať a implementovať pevný etický rámec, bez ktorého by sa rozmachom armádnych systémov AI otvorila Pandorina skrinka.

  Na základe doterajších skúseností v oblasti limitov a rizík súčasných systémov AI je však otázne, či v celom spektre technológií a algoritmov umelej inteligencie bude možné pevný etický rámec dodržať. Vieme si to predstaviť pri cielenom – čo môže znamenať aj limitujúcom a tým aj znevýhodňujúcom – dizajne systémov AI s dôrazom na dodržiavanie navrhnutého rámca (ethics by design), avšak pri súčasnej snahe o čo najširšie adoptovanie umelej inteligencie v armáde a získanie konkurenčnej výhody sme v tomto smere pomerne skeptickí.

402 Kapitola 2.7.8.

- [403]Podľa štúdie *Humans rely more on algorithms than social influence as a task becomes more difficult*, pri ktorej sme sa záverom druhej kapitoly pristavili, sme náchylní týmto systémom umelej inteligencie a priori dôverovať za podmienky, že ich prejavy a výstupy spadajú do rámca exaktnosti, logiky a matematickej presnosti, t.j. do intelektuálneho rámca, ktorý sme si podvedome vytvorili, resp. prijali pre vedu a techniku.

  A priori dôvera – nedbajúc na limity a riziká technológií AI – môže byť veľmi nebezpečným trendom a v určitých rozmeroch (psychologický a sociologický vplyv, algokracia, vojenské nasadenie) aj ohrozením fungovania modernej spoločnosti.

  Preto v kontexte rastúcej dôvery v systémy umelej inteligencie a vnímajúc problémy, s ktorými sa nasadenie týchto systémov potýka i oblastí, ktorých sa ich využitie dotýka, môžeme a musíme stanoviť podmienky, bez splnenia ktorých by nasadenie systémov AI do reálneho sveta, v ktorom interagujú s človekom a vplývajú na spoločnosť, nemalo byť umožnené.

Uvedený sumár problémov technológií umelej inteligencie s priamym či nepriamym dopadom na človeka a spoločnosť nás núti **vytvoriť si všeobecné zásady prístupu k fenoménu umelej inteligencie**:

- nutnou podmienkou vývoja a reálneho nasadenia akéhokoľvek systému AI musí byť **schopnosť a možnosť človeka prebrať kedykoľvek kontrolu nad týmto systémom, resp. právo a možnosť verifikovať a prehodnotiť výsledky jeho činnosti**.
- dizajn a funkcionalita každého systému AI, ktorý má byť reálne nasadený a využívaný, musí byť dostatočne známa, pochopiteľná a verifikovateľná. Jednoducho **musíme vedieť, ako daný systém AI funguje**.
- musíme byť schopní **analyzovať a chápať dôsledky nasadenia konkrétnej technológie AI**.
- **potrebujeme mať jasný manažment rizík**, ktorý zahŕňa riešenie zlyhania a limitných situácií činnosti algoritmov AI.
- musíme **nastaviť primerané požiadavky a podmienky (regulácie) pre dizajn, fungovanie a využívanie systémov AI, ktoré vychádzajú z etických princípov,**

403 Kapitola 2.8.

právnych noriem a morálnych hodnôt.

- nasadenie akéhokoľvek systému AI, ktorý má potenciál ovplyvniť fungovanie, bezpečnosť i budúcnosť celej spoločnosti, **musí byť pod verejným dohľadom a kontrolou**.
- akákoľvek technológia umelej inteligencie **musí byť dôveryhodná, spĺňajúc požadované požiadavky legislatívne, etické i bezpečnostné a zameraná na dobro človeka**.

V prehľade etických výziev, prameniacich z limitov a rizík súčasných systémov umelej inteligencie, ktoré sme uviedli v 2. kapitole i diskutabilných spôsobov a oblastí využitia, prípadne zneužitia, tak nachádzame **tri oblasti nutnej implementácie etických noriem, eticko-právnych regulácií a morálnych zásad**:

- etické normy, zákonné regulácie a morálne zásady **tvorcov** systémov AI
- etické normy, zákonné regulácie a morálne zásady **poskytovateľov a používateľov** týchto systémov
- implementované eticko-právne požiadavky a obmedzenia priamo **v systémoch AI**

Uvedené úvahy primárne vzťahujeme na systémy ANI[404], u ktorých z podstaty veci musíme rátať so zaangažovaním ľudského faktora vo všetkých oblastiach tvorby, používania i realizácie prostriedkov umelej inteligencie. Preto **aplikovanie etických princípov a regulácií v ANI – akokoľvek náročným až neuskutočniteľným by sa to v praxi mohlo ukázať – má vďaka prítomnosti ľudského faktora jasné zásady, vyhranené oblasti a viac menej definované kritériá**.

Takmer všetkými odborníkmi v oblasti kybernetickej bezpečnosti je za najslabší článok bezpečnostného reťazca považovaný človek, ktorý s informačnými systémami interaguje. V oblasti aplikácie eticko-právnych regulácií v systémoch umelej inteligencie sme si osobitne v rámci 2. kapitoly ukázali, že pri týchto technológiách to tak nemusí byť. Keďže nie je prakticky možné pokryť všetky etické výzvy spojené so systémami AI pomocou technologických riešení – **treba investovať nemalé úsilie do vzdelania, osvety**

404 Len pripomeňme, že hovoríme o úzko špecializovaných systémoch umelej inteligencie (narrow AI), ktoré sú optimalizované na zvládnutie konkrétnej úlohy, resp. množiny úloh. Ide súčasne o systémy slabej umelej inteligencie (weak AI), ktoré vykazujú inteligentné správanie na základe modelov a aplikovaných metód i tréningových dát. Sú to teda systémy zamerané na riešenie konkrétnych úloh a sú závislé na ľudskom vstupe a konfigurácii.

i prevencie a formovať morálne postoje vývojárov, poskytovateľov i používateľov týchto technológií.

Edukácia a osveta sa tak stávajú nutnou podmienkou pre rast spoločenskej citlivosti a zodpovednosti v oblasti umelej inteligencie, a to vo viacerých oblastiach:

- spoločenská osveta a vzdelávanie za účelom rastu všeobecného povedomia o limitoch, rizikách a reálnych možnostiach technológií umelej inteligencie.[405]
- trh s umelou inteligenciou v súčasnosti zažíva extrémny boom a vyžaduje veľa špecialistov v tejto oblasti. Utešene nám rastú počty robotníkov AI (riešia nasadenie a prevádzku bežných systémov AI)[406], no nie tak počty odborníkov na AI, ktorí „vidia do hĺbky“ týchto systémov a dokážu riešiť nielen technologické, ale aj eticko-právne výzvy pre zavádzanie dôveryhodných technológií umelej inteligencie.
- tímy odborníkov na AI by mali mať interdisciplinárny rozmer, bez ktorého ťažko uchopiť problematiku dôveryhodnosti umelej inteligencie a jej zamerania na človeka.

Smerovanie k AGI[407], teda k všeobecnej a silnej umelej inteligencii, nás však aj v oblasti etiky stavia pred nové výzvy a takmer určite ešte neznáme možnosti, riziká i obmedzenia.

Ponajprv nie je jasné, kde sa nachádza, resp. bude nachádzať hranica medzi ANI a AGI.

---

405 Jednou z výborných aktivít, ktorá sa rozšírila prakticky v celej Európskej únii, je vytvorenie kurzu Elements of AI z dielne Helsinskej univerzity a organizácie Reaktor Education. Ide o kurz základov umelej inteligencie, ktorý prináša elementárny vzhľad do problematiky AI a poukazuje na viaceré etické výzvy, ktoré sú s umelou inteligenciou spojené. Doteraz kurz absolvovalo cez 750 tisíc občanov EÚ. *Elements of AI*. [on-line]. [cit. 30. marca 2022].
Dostupné na internete: <https://www.elementsofai.com/>

406 V zásade sú nahraditeľní prostredníctvom budúcich sofistikovanejších systémov AI:-)
Problematiku robotníkov vs. odborníkov AI diskutoval prof. Peter Sinčák z Technickej univerzity v Košiciach na konferencii ITAPA: Umelá inteligencia v bežnom živote.
Umelá inteligencia v bežnom živote. [on-line]. Bratislava: ITAPA, 2021. [cit. 24. júna 2021].
Dostupné na internete: <https://www.itapa.sk/12826-sk/program/>

407 Skutočná umelá inteligencia, ktorá je všeobecná (general) a silná (strong). Všeobecná, keďže dokáže zvládnuť akúkoľvek intelektuálnu úlohu a má schopnosť generalizovať, t. j. zovšeobecňovať a prenášať, či adaptovať naučené schopnosti na iné úlohy. Silná, pretože aj skutočne rozumie tomu, či rieši a vykonáva.

Tiež nevieme, do akej miery bude táto hranica jasná a presne definovaná. Je možné, že nás prekvapí **existencia „čiastočných AGI“, resp. vysoko sofistikovaných systémov na pomedzí ANI a AGI**, čo si bude vyžadovať hľadanie svojich vlastných riešení etiky a regulácie.

Ak prezumujeme úspešný vývoj skutočne mysliacej umelej inteligencie a na základe súčasných skúseností z vývoja systémov hlbokého učenia predpokladáme, že táto uvedomelá inteligencia bude inšpirovaná činnosťou ľudského mozgu, otvára sa nám úplne nová perspektíva spôsobu riešenia etických výziev a regulácií smerujúca viac k psychológii, kognitívnym vedám a právu než k informatike, kybernetike a dátovej vede.

V prípade „skutočnej“ umelej inteligencie inšpirovanej ľudským mozgom sa v oblasti etiky a regulácie zmysluplne nepohneme ďalej, pokiaľ nebudeme mať vyriešený jej model správania sa na spôsob teórie mysle a schopnosti zdravého rozumu u človeka.[408]

V mnohých diskusiách o možnostiach a schopnostiach systémov umelej inteligencie dodržiavať eticko-právne regulácie **narážame na nepochopenie inakosti „inteligencie“ týchto technológií**. Uvedomujúc si bytostné rozdiely medzi človekom a systémami AI sa nemôžeme stavať k týmto systémom ako k niečomu ľudsky inteligentnému a človeku podobnému.

Jednu z implikácií, ktoré sa od tohoto omylu odvíjajú, sme už spomenuli v diskusiách o reguláciách LAWs – išlo o tzv. „následnú reguláciu“[409], ktorá nevychádza zo všeobecne akceptovaných mravných hodnôt a etických princípov, ale sa prispôsobuje vývoju a možnostiam nasadenia týchto autonómnych zbraňových systémov. **V podstate toto prispôsobovanie a vývoj v oblasti etiky, hodnotového rámca a morálnej obhájiteľnosti predpokladá u systémov AI podobný model správania, ako je ten ľudský a tým aj možnosť nahradiť ľudský morálny úsudok a svedomie algoritmickým spracovaním**. Znovu pripomíname – pokiaľ nemáme vyriešený model správania sa systémov AI na spôsob teórie mysle a schopnosti zdravého rozumu u človeka – **ide o veľmi nebezpečný argumentačný faul**.

Podobným argumentačným zlyhaním je i porovnávanie presnosti činnosti bojových

408 Modely správania na spôsob teórie mysle a schopnosti zdravého rozumu u človeka sme aspoň okrajovo diskutovali v kapitole 2.1.2.

409 „Následnú reguláciu“ sme rozoberali na konci kapitoly 2.7.8.

systémov AI (osobitne budúcich AGI) s konaním ľudských osôb nasadených v porovnateľných bojových situáciách. Vraj na základe „vhodného naprogramovania“ (po kapitolách 1 a 2 si aspoň približne vieme predstaviť, resp. si ani nedokážeme predstaviť, čo všetko by také niečo u AGI obnášalo) budú LAWs schopné lepšie dodržiavať humanitárne štandardy než ľudia. Určite môžeme súhlasiť, že špičkové LAWs dokážu presnejšie strieľať, rýchlejšie identifikovať jednotlivé objekty scény a možno i bleskurýchle vyhodnocovať operačnú situáciu, atď. – a to všetko bez únavy či iných podmienok zabezpečujúcich životné funkcie ľudského personálu. Takže systémy AI sa budú vedieť niektorých chýb lepšie vyvarovať než ľudia. Na tejto istej – povedzme „technologickej“ úrovni – však môžeme poukázať aj na zlyhania prameniace z rizík a problémov uvedených v kapitolách 2.1. až 2.4., takže z iného uhla pohľadu by mohli mať navrch ľudia.[410] Problematická je však iná, hlbšia rovina. V kontexte modelu ľudského správania poznajúc ľudské zlyhania, ktoré môžu viesť až k vojnovým zverstvám, sa snažíme nastaviť záväzné pravidlá a dohovory i trestno-právne postihy na obmedzenie týchto zlyhaní.[411] **Pokiaľ však nemáme vyriešený model správania sa AGI, jej základné myšlienkové rámce korešpondujúce u človeka so zdravým rozumom, nemáme na základe čoho a ako nastavovať pravidlá, ktoré by boli vymožiteľné a aj reálne dodržiavané**.

V niektorých oblastiach sa môžeme dostať až k principiálnej neistote v oblasti etických záruk. Napríklad, ak znovu uvažujeme o súčasných autonómnych zbraňových systémoch so smrtiacim účinok (LAWs) ako o systémoch ANI, správne môžeme argumentovať, že pokiaľ tieto systémy nedokážu chápať bojovú scénu a na ňu naviazané široké súvislosti (napríklad rozpoloženie identifikovaných ľudských subjektov, výskyt civilistov na bojisku, použitie civilistov ako štít a pod.), nemôžu bez riadiaceho ľudského činiteľa konať. Ak však by sme mali LAWs vybavený algoritmami AGI, pokiaľ si nebudeme istí ich modelom správania porovnateľným s ľudskou schopnosťou zdravého rozumu a tým pádom aj zodpovedajúco formovateľným, o to skôr musíme povedať, že takéto systémy nemôžu samostatne konať!

Ako sme uviedli v kapitole 2.1.2., človek – napriek tomu, že neustále robí chyby – dokáže konať skutočne inteligentne. Jednoducho má to, čo chýba všetkým súčasným systémom

410 Samozrejme tým nerozporujeme vojenské benefity, ktoré sme uviedli v kapitole 2.7.

411 Napríklad Ženevské konvencie, ktoré upravujú podmienky a pravidlá medzinárodného práva na ochranu obetí vojny alebo Haagske konvencie a Ženevský protokol, ktoré upravujú použitie zbraní vo vojne.

umelej inteligencie, a to zdravý rozum (ako súčasť všeobecnej inteligencie), ktorého súčasťou je schopnosť abstrahovať a na základe analógií a konceptov nachádzať riešenia, myslieť a riešiť na základe nesmierneho rozsahu najrozmanitejších vedomostí, poznatkov a skúseností. Ľudská bytosť využíva zdravý rozum podvedome a spontánne v akejkoľvek oblasti života. Preto pre mnohých je dôveryhodný systém umelej inteligencie, ktorý môže úspešne a samostatne fungovať v komplexnom reálnom svete, podmienený schopnosťou mať zdravý rozum tak, ako má človek.

Tieto úvahy nám môžu poslúžiť aj **v diskusii o schopnosti implementovať eticko-právne požiadavky a obmedzenia priamo v systémoch AI**.

Vieme si predstaviť parciálnu implementáciu zameranú napríklad na odolnosť voči predsudkom a poverám algoritmov ANI, no nie komplexné zvládnutie etických princípov a právnych noriem. I v oblasti etiky sa tak pre nás stáva lákadlom „uvedomelá“ AGI, schopná skutočne konať inteligentne, pre ktorú by komplexnosť etických princípov nemusela byť problémom. Ako sme však už uviedli, pokiaľ nebudeme mať vyriešený model správania so schopnosťou prejavovať (a zachovávať) zdravý rozum, s plnou implementáciou eticko-právnych požiadaviek a obmedzení nemôžeme rátať.

V komplexnom pohľade na človeka si môžeme uvedomiť viaceré rozmery mantinelov a obmedzení, ktoré sa podpisujú pod naše schopnosti určitým spôsobom konať:

1. ako biologické systémy staviame na tom, čo do DNA vložila evolúcia a preverila príroda, pričom presah do kognitívnych schopností a dopad na psychológiu človeka netreba v dnešnej dobe pripomínať.
2. využívame zdravý rozum (ako súčasť všeobecnej inteligencie), ktorého súčasťou je schopnosť abstrahovať a na základe analógií a konceptov nachádzať riešenia, myslieť a riešiť využívajúc bázu nesmierneho rozsahu najrozmanitejších vedomostí, poznatkov a skúseností.
3. máme hodnotový systém a využívame sociologické i kultúrne rámce vnímania sveta a konania.
4. máme svedomie a duchovný rozmer nášho života.

Uvažujúc o umelej inteligencii, k 1. bodu by sme vedeli nájsť aspoň hrubú analógiu v tvorbe dnešných systémov AI.

Vieme vytvoriť – a aj sme vytvorili – komplexné systémy ANI, ktoré podľa 2. bodu dokážu

riešiť svoje úlohy využívajúc bázu nesmierneho rozsahu najrozmanitejších vedomostí a poznatkov. Neviem však realizovať AGI so schopnosťou plnej abstrakcie za skutočného chápania vzťahov či súvislostí a zdravého rozumu (ako detailne rozoberáme v 5. kapitole).

Dokážeme prevádzkovať technológie ANI, ktoré vedia spracúvať naše sociologické a kultúrne rámce. Sme však na rozpakoch, ako realizovať AGI s implementovaným (nebodaj aj chápaným) hodnotovým systémom.

Svedomie, ako referenčný bod pre etický systém a morálne rozhodovanie, už z princípu nie je možné v ANI „implementovať“. Keďže svedomie nedokážeme oddeliť od osoby so schopnosťou rozumu, slobodnej vôle a sebauvedomenia, v prípade AGI sa potýkame s dilemou umelej inteligencie, u ktorej si nevieme predstaviť ontologickú analógiu ľudského bytia, bez ktorej niet ani svedomia.[412]

Vráťme sa však späť do kráľovstva dnešných technológií umelej inteligencie, teda k úzko špecializovaným systémom slabej umelej inteligencie (ANI) a v nasledovných kapitolách sa aspoň v krátkosti pozrime na to, ako sa súčasný svet snaží uchopiť jej etické problémy a výzvy.

3.2. Angažovanosť a aktivity na poli etiky umelej inteligencie

Jednou z odlišností, ktorými sa stať o systémoch AI vo vojenskej sfére líšila od ostatných častí 2. kapitoly popisujúcich limity a riziká týchto technológií, bolo viacnásobné zdôraznenie angažovanosti a aktivít v oblasti etiky a regulácie nasadenia autonómnych zbraňových systémov.

Dôvod je jasný – prakticky žiadna iná aplikácia systémov umelej inteligencie nepriniesla toľko polemík a rizík ako práve vojenské využitie. Obavy z napredovania vo vývoji algoritmov umelej inteligencie a spôsobu ich zavádzania do moderných zbraňových systémov tak dali v poslednej dekáde vznik viacerým iniciatívam, či už odborným a celo-spoločenským alebo armádnym.[413]

Na chvíľku prekročiac rámec posledného desaťročia, pokladáme za vhodné spomenúť jednu zo základných výziev Russell-Einsteinovho manifestu, reagujúceho na masívny rozmach vývoja a zavádzania nukleárnych zbraní do výzbroje veľmocí studenou vojnou

412 A ak by sme si to i vedeli predstaviť, ako by v takomto prípade vyzeral teonómny rozmer svedomia, ktoré má referenčný bod a autoritu mimo seba v Bohu?

413 Niektoré sme už v kapitole 2.7.8. spomínali, no dovolíme si ich predsa len zopakovať.

rozdeleného sveta:[414]

„Mnohé varovania vyslovili významní muži vedy a autority v oblasti vojenskej stratégie. Nikto z nich nepovie, že najhoršie výsledky sú isté. Hovoria však, že tieto výsledky sú možné a nikto si nemôže byť istý, že sa neuskutočnia. Zatiaľ sme nezistili, že by názory odborníkov na túto otázku v nejakej miere záviseli od ich politického presvedčenia alebo predsudkov. Pokiaľ to naše výskumy odhalili, závisia len od rozsahu vedomostí konkrétneho odborníka. Zistili sme, že ľudia, ktorí toho vedia najviac, sú najpochmúrnejší.“

V tomto kontexte je skutočne nadčasová najznámejšia a pritom tak strohá výzva manifestu: „**Pamätajte na svoju ľudskosť a zabudnite na zvyšok**“.

Už v roku 2013 skupina inžinierov, odborníkov na umelú inteligenciu a robotiku a ďalších vedcov a výskumníkov z tridsiatich siedmich krajín sveta vydala **Scientists' Call to Ban Autonomous Lethal Robots**. V tejto výzve sa uvádza, že chýbajú vedecké dôkazy o tom, že by roboty mohli v budúcnosti disponovať „funkciami potrebnými na presnú identifikáciu cieľa, situačné povedomie alebo rozhodnutia týkajúce sa primeraného použitia sily“. LAWs tak môžu spôsobiť vysokú mieru vedľajších škôd. Vyhlásenie sa končí zdôraznením, že „rozhodnutia o použití hrubej sily sa nesmú delegovať na stroje“.[415]

V júli roku 2015 bol na jednej z medzinárodných konferencií o umelej inteligencii zverejnený otvorený list **Autonomous Weapons: An Open Letter from AI & Robotics Researchers**, ktorý vyzýval na zákaz autonómnych zbraní. V liste sa priamo píše: „technológia umelej inteligencie dosiahla bod, v ktorom je nasadenie týchto systémov

414 Russell-Einsteinov manifest, vydaný v roku 1955, bol verejným vyhlásením významných vedcov v čase zvýšeného medzinárodného napätia medzi Východom a Západom, ako aj rastúceho zasahovania do vnútorných záležitostí štátov prostredníctvom vojenských i tajných operácií. Bolo to tiež obdobie, ktoré charakterizoval prevratný vedecký pokrok a technologické inovácie.
Manifest mal pôvod v obavách fyzika Josepha Rotblata a polyhistora Bertranda Russella z účinkov jadrových zbraní, rizika šírenia jadrových zbraní a budúcnosti ľudstva. To znamená, že autori chceli biť na poplach v súvislosti s potenciálne katastrofickými dôsledkami použitia jadrových zbraní a existenčným rizikom, ktoré pre ľudstvo predstavuje ich zachovanie. Medzi jedenástimi signatármi manifestu bolo niekoľko najvýznamnejších svetových vedcov. Desať z nich bolo nositeľmi Nobelovej ceny za prínos v oblasti fyziky, chémie, fyziológie a medicíny, literatúry alebo mieru.
VIGNARD, K. *Manifestos and open letters: Back to the future?* [on-line]. [cit. 21. marca 2022]. Dostupné na internete: <https://thebulletin.org/2018/04/manifestos-and-open-letters-back-to-the-future/>

415 *Scientists' Call to Ban Autonomous Lethal Robots*. ICRAC, October 2013. [on-line]. [cit. 8. marca 2022]. Dostupné na internete: <http://www.icrac.net/>

prakticky, ak nie legálne, uskutočniteľné v priebehu nie desaťročí, ale rokov a v stávke je veľa: autonómne zbrane sa uvádzajú ako tretia revolúcia v zbraňových systémoch, po strelnom prachu a jadrových zbraniach“. Tento otvorený list poukazuje na pokrok i výhody spojené s vývojom systémov AI, zvažuje riziká a dopady zneužitia LAWs na celú oblasť umelej inteligencie a graduje k výzve na „zákaz útočných autonómnych zbraní, ktoré sú mimo zmysluplnej ľudskej kontroly“.[416]

List mal do roku 2016 pôsobivý zoznam signatárov, medzi ktorými sú priekopníci z oblasti vedy a techniky[417], takmer päťtisíc výskumníkov v oblasti umelej inteligencie a robotiky i skoro dvadsaťsedem tisíc ďalších podporovateľov.

Nasledovali i cielenejšie aktivity, napríklad otvorený list výskumníkov a zakladateľov spoločností pôsobiacich v oblasti umelej inteligencie **An Open Letter to the United Nations: Convention on Certain Conventional Weapons**, v ktorom v roku 2017 vyzývali všetkých participantov v tejto oblasti vývoja, aby „zabránili pretekom v zbrojení týmito zbraňami, aby chránili civilistov pred ich zneužitím a aby zabránili destabilizujúcim účinkom týchto technológií“.[418]

Otvorený list nenechal nikoho na pochybách, že signatári majú vážne obavy z uplatňovania technológií umelej inteligencie a robotiky v budúcich zbraňových systémoch: „Nemáme veľa času na to, aby sme konali. Keď sa táto Pandorina skrinka otvorí, bude ťažké ju zavrieť.“[419]

Tieto aktivity nezostali bez odozvy – v nasledujúcich rokoch viacero organizácií podieľajúcich sa na vývoji algoritmov umelej inteligencie aplikovateľných vo vojenských systémoch, resp. v akýchkoľvek systémoch AI zneužiteľných v armáde a represívnych

---

416 *Autonomous Weapons: An Open Letter from AI [Artificial Intelligence] & Robotics Researchers*. [on-line]. Future of Life Institute, 2015. [cit. 8. marca 2022].
Dostupné na internete: <http://futureoflife.org/open-letter-autonomous-weapons/>

417 Okrem iných napríklad Elon Musk (vynálezca a zakladateľ spoločnosti Tesla), Steve Wozniak (spoluzakladateľ spoločnosti Apple), fyzik Stephen Hawking (Univerzita v Cambridge), Noam Chomsky (Massachusettský technologický inštitút), Stuart Russell (Berkeley) atď.

418 *An Open Letter to the United Nations: Convention on Certain Conventional Weapons.* [on-line]. [cit. 21. marca 2022].
Dostupné na internete: <https://futureoflife.org/2017/08/20/autonomous-weapons-open-letter-2017/>

419 VIGNARD, *Manifestos and open letters: Back to the future?* [on-line]. [cit. 21. marca 2022].
Dostupné na internete: <https://thebulletin.org/2018/04/manifestos-and-open-letters-back-to-the-future/>

zložkách, odrieklo svoju účasť na projektoch z týchto oblastí.[420] Paradoxom však je, že viaceré z týchto organizácií odmietlo obmedziť svoje technológie umelej inteligencie v tak kontroverznej oblasti ako je kapitalizmus dohľadu (surveillance capitalism), ktorý sme diskutovali v kapitole 2.5. a ktorého dôsledky sú v niektorých aspektoch pre spoločnosť takmer tak rizikové ako autonómne zbraňové systémy.[421]

Uvádzajúc aktivity v oblasti etiky autonómnych zbraňových systémov treba pripomenúť aj armádne aktivity, ktorých existencia, rozsah a praktický dopad majú svoju výpovednú hodnotu a dotvárajú hodnotový rámec tej-ktorej armády. Či už ide o univerzitné armádne aktivity v USA, z prostredia ktorých sme v kapitolách 2.7.7. a 2.7.8. citovali z článku *Pros and Cons of Autonomous Weapons Systems,* alebo priamo o eticko-právny diskurz prebiehajúci v armádnych kruhoch, napríklad v kapitole 2.7.8. komentovaný dokument **AI Principles: Recommendations on the Ethical Use of Artificial Intelligence by the Department of Defense** z Defense Innovation Board USA.

Ťahúňom armádnych aktivít na poli regulácií a eticko-právnych dôsledkov nasadenia LAWs sú primárne armády USA a NATO, nakoľko odzrkadľujú nielen spoločenské povedomie v tejto oblasti, ale predovšetkým technologickú vyspelosť armádnych systémov AI. Podľa viacerých autorov im technologicky nielen zdatne sekunduje, ale ich možno aj predbieha rozvoj čínskych zbraňových systémov a všeobecne technológií AI,[422] čo však – vzhľadom na súčasný postoj čínskych úradov k etike systémov AI – môže byť problémom pre celosvetové presadzovanie potrebných regulácií v tejto oblasti. A vzhľadom na zavádzanie technológií AI aj do iných armád disponujúcich zbraňami hromadného ničenia, ktoré môžu mať skutočne rôznorodý vzťah k eticko-právnym aspektom AI,

420 SAMPLE, I. *Thousands of leading AI researchers sign pledge against killer robots*. [on-line]. [cit. 21. marca 2022].
Dostupné na internete: <https://www.theguardian.com/science/2018/jul/18/thousands-of-scientists-pledge-not-to-help-build-killer-ai-robots>

421 ZUBOFF, SH. *The real reason why Facebook and Google won't change*. [on-line]. [cit. 21. marca 2022].
Dostupné na internete: <https://www.theguardian.com/science/2018/jul/18/thousands-of-scientists-pledge-not-to-help-build-killer-ai-robots>

422 *The AI arms race*. [on-line]. [cit. 21. marca 2022].
Dostupné na internete: <https://www.ft.com/content/21eb5996-89a3-11e8-bf9e-8771d5404543>
'Wake up': Ex-Air Force software officer warns China is winning AI battle. [on-line]. [cit. 21. marca 2022].
Dostupné na internete: <https://www.washingtontimes.com/news/2021/oct/11/nicolas-chaillan-warns-china-winning-ai-battle/>

musíme brať vážne výzvy na seriózne riešenie tejto problematiky v snahe vyvarovať sa katastrofe nukleárnej vojny.[423]

Angažovanosť výskumníkov a odbornej verejnosti v oblasti regulácie zbraňových systémov umelej inteligencie našla odozvu aj v diplomatických diskusiách OSN v Ženeve, ktoré sa vedú v rámci Dohovoru o určitých konvenčných zbraniach z roku 1980. Konzultácie, ktoré sa započali v roku 2014 ako séria neformálnych stretnutí, v roku 2017 vykryštalizovali do stretnutí skupiny vládnych expertov. I napriek tomu, že členovia vedeckej komunity a experti sú opatrní a hľadajú riešenia potenciálnych nebezpečenstiev autonómnych zbraní, určite nie sú pesimistickí, pokiaľ ide o potenciálne výhody umelej inteligencie alebo technologických inovácií vo všeobecnosti. Sú preto aktívne zaangažovaní v úsilí, aby umelá inteligencia bola prospešná pre spoločnosť a v konečnom dôsledku aj pre ľudstvo.[424]

Jedným z plodov týchto aktivít v rámci Európskej únie bolo prijatie **Uznesenia o autonómnych zbraňových systémoch**[425], ktoré sa v roku 2018 uskutočnilo na pôde Parlamentu EÚ a ktoré si kladie za cieľ globálny zákaz smrtiacich autonómnych zbraňových systémov. Asi nás neprekvapí odmietavý postoj USA a Ruska (ale i ďalších dôležitých hráčov v oblasti LAWs) k tomuto zákazu. Ako sme uviedli – vzhľadom na potenciál LAWs pracujú skôr na reguláciách a „korektnom“ spôsobe ich nasadenia.

Vzhľadom na smrtiace účinky a rozsah rizík je pochopiteľné, že práve táto oblasť je v strede záujmu, no nie je jediná.

V kapitolách 2.5. a 3.1. sme spomenuli aj problematiku autonómnych vozidiel, v rámci ktorej sa rieši nielen schopnosť autonómnych systémov správne sa rozhodovať v kritických situáciách, ale i legislatívne požiadavky na tzv. bezpečnostného operátora (safety operator), hľadajúc objektívnu mieru zodpovednosti ľudského faktora pri riadení autonómneho vozidla v rámci jednotlivých stupňov automatizácie, ktoré sme uviedli

423 GEIST, E., LOHN, A. J. *How Might Artificial Intelligence Affect the Risk of Nuclear War?* [on-line]. Santa Monica, CA: RAND Corporation, 2018. [cit. 21. marca 2022].
Dostupné na internete: <https://www.rand.org/pubs/perspectives/PE296.html>

424 VIGNARD, *Manifestos and open letters: Back to the future?* [on-line]. [cit. 21. marca 2022].
Dostupné na internete: <https://thebulletin.org/2018/04/manifestos-and-open-letters-back-to-the-future/>

425 *Resolution on autonomous weapon systems*. [on-line]. [cit. 21. marca 2022].
Dostupné na internete: <https://oeil.secure.europarl.europa.eu/oeil/popups/ficheprocedure.do?reference=2018/2752(RSP)&l=en>

v kapitole 2.5.

Asi najďalej sú v tomto smere diskusie v USA, ktoré sa premietli do viacerých legislatívnych ustanovení, či už na úrovni samostatných štátov (napr. Kalifornia, Texas) alebo v rámci federálnych zákonov. V súčasnosti (marec 2022) napríklad Spojené štáty menia svoju legislatívu, z ktorej pre autonómne vozidlá 5. úrovne autonómie vypadla povinnosť zahrnúť do ich vybavenia základné ovládacie prvky.[426] **I v tejto oblasti sme viackrát deklarovali požiadavku na schopnosť človeka kedykoľvek prebrať kontrolu nad systémom. Ako vidíme, čím sofistikovanejší systém umelej inteligencie v autonómnych vozidlách sa použije, tým viac sa naplnenie tejto požiadavky rozplýva. Ide o nebezpečný trend, ktorý by sa mohol rozšíriť i na iné oblasti nasadenia pokročilých systémov AI**.

Problematika autonómnych vozidiel je na eticko-právne výzvy veľmi bohatá – či už ide o spoľahlivú funkcionalitu systémov AI za každých poveternostných podmienok a možných situácií v bežnej cestnej prevádzke, alebo o správne rozhodovanie sa v kritických situáciách[427] a v neposlednom rade aj o ochranu osobných údajov a spoločenské dopady, keďže prevádzka týchto vozidiel je extrémne závislá na širokospektrálnom zbere najrozličnejších dát a ich spracovaní v reálnom čase algoritmami AI, resp. i následnom analytickom vyťažení v dátových centrách poháňaných technológiami umelej inteligencie.

V oblasti dohľadových systémov a hlavne komplexnej problematiky algoritmického riadenia míľovými krokmi kráča v ústrety umelej inteligencii Čína.[428] Je však len na škodu, že tento prekotný rozvoj nie je sprevádzaný adekvátnou verejnou diskusiou a nepodlieha odbornému i spoločenskému diskurzu, ktorý by bol garantom, resp. katalyzátorom eticko-právnych požiadaviek sprevádzajúcich tento pod taktovkou vládnej ideológie prebiehajúci

426 SHEPARDSON, D. *U.S. eliminates human controls requirement for fully automated vehicles*. [on-line]. [cit. 22. marca 2022].
Dostupné na internete: <https://www.reuters.com/business/autos-transportation/us-eliminates-human-controls-requirement-fully-automated-vehicles-2022-03-11/>

427 Už sme spomínali príklad kritickej situácie, v ktorom sa musí autopilot rozhodnúť medzi potencionálnym ohrozením života posádky alebo ostatných účastníkov v čase blížiacej sa alebo prebiehajúcej dopravnej nehody.

428 Pomerne ucelený pohľad na problematiku zavádzania algokracie v Číne ponúka súbor štúdií britskej agentúry pre sociálny rozvoj Nesta.
The AI Powered State. [on-line]. [cit. 26. marca 2022].
Dostupné na internete: <https://www.nesta.org.uk/feature/ai-powered-state/>

technologický šprint.[429]

Vo všeobecnosti môžeme povedať, že povedomie o etických požiadavkách na systémy umelej inteligencie neustále rastie. Riziká technológií AI riešia nielen expertné skupiny OSN, EÚ, či jednotlivých štátov, ale varujú pred nimi i viaceré osobnosti známe z vedeckého a technologického sveta (Hawking, Musk,...). A prakticky pri všetkých významnejších centrách výskumu umelej inteligencie sa etablujú oddelenia zaoberajúce sa etickými otázkami a výzvami.

Mnohé z iniciatív, ktoré prebiehajú vo vládnych a vedeckých pracovných skupinách i technologických komunitách, sa skutočne snažia formulovať potenciálne prínosy umelej inteligencie a súčasne obmedzovať jej riziká. Medzi významné výsledky týchto snáh môžeme zarátať vedecký program formulovaný v **Otvorenom liste o výskumných prioritách pre robustnú a prospešnú umelú inteligenciu**[430] z roku 2015 a **Globálnu iniciatívu IEEE o etike autonómnych a inteligentných systémov**[431]. Medzi ďalšie iniciatívy na multilaterálnej úrovni patrí napríklad AI for Good Global Summit 2017[432], ako aj práca organizácií ako OpenAI[433] alebo Partnership on AI[434], na ktoré nadväzuje mnoho ďalších aktivít po celom svete.

---

429 Hodnotíme tak podľa našich kritérií euroatlantického spoločenského priestoru. Aktivity na podporu implementácie etických pravidiel samozrejme prebiehajú aj v Číne, napr.:
Understanding China's AI Strategy. [on-line]. [cit. 21. marca 2022].
Dostupné na internete: <https://www.cnas.org/publications/reports/understanding-chinas-ai-strategy>
China wants to make its own rules for AI ethics. [on-line]. [cit. 21. marca 2022].
Dostupné na internete: <https://www.scmp.com/abacus/tech/article/3029194/china-wants-make-its-own-rules-ai-ethics>

430 *Open Letter on Research Priorities for Robust and Beneficial Artificial Intelligence*. [on-line]. [cit. 21. marca 2022].
Dostupné na internete: <https://futureoflife.org/2015/10/27/ai-open-letter/>

431 *IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems*. [on-line]. [cit. 21. marca 2022].
Dostupné na internete: <http://standards.ieee.org/develop/indconn/ec/ead_v2.pdf>

432 *AI for Good Global Summit 2017*. [on-line]. [cit. 21. marca 2022].
Dostupné na internete: <https://www.itu.int/en/ITU-T/AI/Pages/201706-default.aspx>

433 *OpenAI*. [on-line]. [cit. 21. marca 2022].
Dostupné na internete: <https://openai.com/>

434 *Partnership on AI*. [on-line]. [cit. 21. marca 2022].
Dostupné na internete: <https://partnershiponai.org/>

3.3. Legislatívne kroky a regulácie

Až donedávna v žiadnej z krajín sveta neexistovala ucelená legislatíva pokrývajúca celú problematiku umelej inteligencie. Doterajšie regulácie dotýkajúce sa technológií AI boli buď veľmi špecifické, riešiac parciálne problémy konkrétnych oblastí nasadenia týchto systémov a / alebo boli súčasťou iných regulácií, napr. kybernetickej bezpečnosti, ochrany osobných údajov, sektorových regulácií v rámci finančného sektora alebo štátnej správy.

V rámci Európskej únie však vzniká úplne nová regulácia, ktorý nemá obdobu nikde vo svete.[435] Stanovujúc si za cieľ pokryť celú problematiku súčasných systémov umelej inteligencie, ide vôbec o prvú komplexnú reguláciu konkrétnej technológie.[436]

Primárnym dôvodom pre vznik a aplikovanie tejto regulácie je dopad fenoménu umelej inteligencie na človeka a spoločnosť. Jednoducho povedané, umelá inteligencia je spôsobilá zmeniť postavenie človeka.

Ľudská bytosť so svojím zmyslovým vnímaním, intelektuálnym vybavením a schopnosťami i komplexnou psychológiou nie je prakticky schopná technológiám tohoto „kalibru“ vôbec odolávať. Spoločnosť tak bez adekvátnych regulácií a nastavených limitov nie je schopná čeliť dôsledkom činnosti systémov a vplyvu technológií AI s potenciálom rozkladať základné princípy, na ktorých je naša spoločnosť postavená, napríklad negatívne vplývať na ľudské práva a dôstojnosť človeka, podrývať demokratické princípy a manipulovať, atakovať bezpečnostné mechanizmy a pod.[437]

---

435 Vo svete v súčasnosti vzniklo viacero regulačných rámcov AI:

– EU AI HLEG Guidelines and Assessment List for Trustworthy AI (z nej pramení európsky Akt o AI)

– UK's ICO AI Auditing Framework

– Singapore Model Governance Framework

– Dubai AI Ethics Toolkit

– Hong Kong Ethical Accountability Framework

Európsky rámec je však podľa analýzy Artificial Intelligence: Principles, laws, and frameworks jednoznačne najvyváženejší, najucelenejší a najkomplexnejší.

Por. *Artificial Intelligence: Principles, laws, and frameworks*. OneTrust DataGuidance Limited, 2022. ISSN 2398-9955.

436 PATTYNOVÁ, J. *Výzvy a právní aspekty umělé inteligence*. In: *Umělá inteligence 2021.* Praha: 2021. Dostupné na internete: <https://www.tuesday.cz/akce/umela-inteligence-1/zaznam-akce/>

437 PATTYNOVÁ, *Výzvy a právní aspekty umělé inteligence*. Dostupné na internete: <https://www.tuesday.cz/akce/umela-inteligence-1/zaznam-akce/>

Tento rozklad môže viesť až k disruptívnym zmenám v modernej informačnej a znalostnej spoločnosti.

Podľa Gartner *Top Strategic Predictions For 2020 And Beyond* „Technológie (primárne AI) a ich aplikácie sú pripravené ovplyvniť každý aspekt toho, čo nazývame človečenstvom“.[438]

Gartner vo svojej podrobnej správe *Top Strategic Predictions for 2020 and Beyond: Technology Changes the Human Condition*[439] odhaduje, že v roku 2023 bude 40% digitálne sledovaného správania realizovaného algoritmami AI a v roku 2050 bude viac než 50% reklám cielených na základe detekcie emócií technológiami umelej inteligencie. Ide o skutočnosti, na ktoré sa ľudská spoločnosť nie je schopná v tak krátkej dobe adaptovať.

Zaujímavou je preto predikcia regulácie, ktorá stavala na existujúcich legislatívnych aktivitách technologicky pokročilých štátov: do roku 2023 štyri z krajín G7 nastavia pravidlá dohľadu nad vývojármi AI a nasadením metód strojového učenia.

Tieto skutočnosti – aspoň rámcovo vysvetľujúce primárny zámer pre vznik a aplikovanie regulácie umelej inteligencie v jej dopade na človeka a spoločnosť – patria ku konkrétnym dôvodom, pre ktoré európski zákonodarcovia prišli s komplexnou reguláciou, o ktorej dúfajú, že podobne ako pri GDPR pôjde o legislatívu ašpirujúcu na globálny štandard, resp. reguláciu zásadným spôsobom ovplyvňujúcu zákonodarstvo aj ostatných krajín.

Univerzálny záber a potenciálny celosvetový dosah v súčasnosti prakticky jedinej komplexnej regulácie technológií umelej inteligencie vo svete je dôvodom, pre ktorý sa v tejto kapitole primárne venujeme len tejto konkrétnej legislatíve.

Okrem Gartnerom spomenutých trendov využívania technológií AI, ktorých riziká sme širšie diskutovali v kapitole 2.5. a ktorých dôsledky ľudská spoločnosť nie je schopná v tak krátkej dobe vstrebať, resp. adaptovať sa na ne, európski zákonodarcovia reagovali i na ďalšie „patológie“ a výzvy:[440]

---

438 *Gartner Top Strategic Predictions For 2020 And Beyond*. [on-line]. [cit. 23. marca 2022]. Dostupné na internete: <https://www.gartner.com/smarterwithgartner/gartner-top-strategic-predictions-for-2020-and-beyond>

439 *Top Strategic Predictions for 2020 and Beyond: Technology Changes the Human Condition*. [on-line]. [cit. 24. marca 2022]. Dostupné na internete: <https://www.gartner.com/document/3970846>

440 PATTYNOVÁ, *Výzvy a právní aspekty umělé inteligence*.

- dilema medzi personalizáciou zákazníckej skúsenosti/reklamy a manipuláciou v dôsledku informačnej asymetrie. V súčasnosti sociálne siete, poskytovatelia služieb, banky a iné organizácie spracúvajú o svojich používateľoch na základe algoritmického spracovania vstupov zákazníkov extrémne veľa (nielen) osobných údajov a sú schopní ďalekosiahleho modelovania a analýzy s následkami, ktoré sme v kapitole 2.5. farbisto rozoberali. Bez regulácie by tento trend nabral obludné rozmery.
- ochrana osôb, spoločnosti a demokracie vzhľadom na riziká deep fake a manipulácií, pri ktorých ide o takú manipuláciu reality (primárne mediálnych záznamov a prenosov, informačných tokov a spravodajstva) prostredníctvom technológií AI, že prakticky nie je možné rozoznať reálne mediálne záznamy a správy od umelých, nemajúcich s realitou nič spoločné. Osobitne obrazový materiál má schopnosť pôsobiť na emocionálnu zložku a prinášať podprahové podnety, voči ktorým – ak sú v podaní sofistikovaných algoritmov AI – môžeme byť takmer bezbranní.
- možnosť zneužitia biometrických technológií a osobitne problematika identifikácie osôb (face recognition) v reálnom čase. V digitalizovanej spoločnosti ide o veľký problém v oblasti dohľadových systémov, sledovania, ochrany súkromia a ľudských práv. Naviac s pomocou biometrických údajov sa v informačnom svete overuje digitálna identita osoby, takže ich zneužitie môže mať fatálne následky z pohľadu práva i etiky.
- riziko sofistikovaných zásahov do súkromia a osobných slobôd z vážnych dôvodov, napr. ochrana zdravia a pod. I keď by sa mohlo zdať, že ide primárne o oblasť, ktorej sme sa venovali pri dohľadových a spravodajských systémoch, čo i len parciálne smerovanie k algoritmickému riadeniu spoločnosti poukazuje na celý rad nových problémov, keďže pre úspešnosť týchto systémov AI je potrebné zasiahnuť do súkromia veľkého množstva ľudí.[441]

Dostupné na internete: <https://www.tuesday.cz/akce/umela-inteligence-1/zaznam-akce/>

441 Napr. pre správne diagnostikovanie vzácnych druhov vážnych chorôb musí algoritmus AI mať prístup k v súčasnosti osobitne chránenej kategórii zdravotných osobných údajov, pre spracúvanie hypoték a úverov v reálnom čase s odolnosťou voči sofistikovaným podvodom musí byť systém AI natrénovaný na nesmiernom množstve finančných dát zákazníkov, podpora pre automatizované súdne konania musí vychádzať z výborne zvládnutých databáz realizovaných súdnych prípadov a pod.

Synergický efekt doterajších parciálnych regulácií, metodického skúmania vplyvu technológií umelej inteligencie na človeka a spoločnosť i rastúceho povedomia a angažovanosti na poli etiky a práva sa na pôde inštitúcií EÚ pretavil do návrhu nariadenia[442] o umelej inteligencii, ktorý pod skráteným názvom **Akt o umelej inteligencii** (Artificial Intelligence Act) bol zverejnený 21. apríla 2021 a aktuálne je v stave revízie a doplnení zo strany Európskeho parlamentu a Rady EÚ.[443]

Nariadenie priamo na pôde EÚ nestavalo na zelenej lúke – už v apríli 2019 prezentovala Skupina expertov AI na vysokej úrovni **Etické usmernenia pre dôveryhodnú umelú inteligenciu**. Išlo o revidované znenie z roku 2018, do ktorého bolo v rámci otvorenej konzultácie zapracovaných viac ako 500 pripomienok.[444] Mimochodom, sumárom usmernení je všeobecná požiadavka na systémy AI, ktoré musia byť:[445]

- zákonné – rešpektujúce všetky platné zákony a predpisy.
- etické – rešpektujúce etické zásady a hodnoty.
- robustné – z technického hľadiska a zároveň zohľadňujúca svoje sociálne prostredie.

Ide o rovnaké závery ako tie, ktoré sme uviedli ku koncu kapitoly 2.8. o podmienkach pre dôveryhodnú umelú inteligenciu.

Nariadenie Akt o umelej inteligencii sa naviac inšpirovalo úspešným zavedením dôležitej

442 Na stupnici nariadenie – smernica – rozhodnutia – odporúčania/stanoviská má nariadenie najväčšiu právnu silu, keďže má prednosť pred vnútroštátnym právom, je všeobecne záväzné pre všetkých, neimplementuje sa – platí v takej forme, v akej ho prijal Európsky parlament a Rada.

443 Plný názov nariadenia znie:
Nariadenie Európskeho parlamentu a Rady, ktorým sa stanovujú harmonizované pravidlá v oblasti umelej inteligencie (Akt o umelej inteligencii) a menia niektoré legislatívne akty únie. [on-line]. [cit. 24. marca 2022].
Dostupné na internete: <https://eur-lex.europa.eu/legal-content/SK/TXT/?uri=CELEX:52021PC0206>

444 *Ethics guidelines for trustworthy AI*. [on-line]. [cit. 28. marca 2022].
Dostupné na internete: <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>

445 V usmerneniach sa zároveň predkladá **sedem kľúčových požiadaviek, ktoré by mali systémy AI spĺňať, aby sa mohli považovať za dôveryhodné: ľudské zastúpenie a dohľad; technická odolnosť a bezpečnosť; ochrana súkromia a správa údajov; transparentnosť; rozmanitosť (diversity), nediskriminácia a spravodlivosť; spoločenský a environmentálny blahobyt; zodpovednosť**.

regulácie v oblasti ochrany osobných údajov, tzv. GDPR[446] a prevzalo z neho základné regulačné schémy, aplikované na problematiku umelej inteligencie a spojené s aktuálnym stavom poznania v oblasti etiky technológií AI.

Akt o umelej inteligencii:

- nadväzujúc na „White Paper“ EÚ bazíruje na *human-centered* prístupe, teda **požaduje také systémy AI, ktoré sú zamerané na človeka**. Bez splnenia tejto podmienky nie je možné hovoriť o dôveryhodnej umelej inteligencii.[447]
- **zavádza vysoký štandard povinností**. Doteraz naakumulované osvedčené postupy (best practices) z oblasti implementácie systémov AI zameraných na človeka, ktoré spĺňajú eticko-právne požiadavky, prenáša do podoby zákona.
- **aplikuje rozšírenú teritoriálnu pôsobnosť**, keďže jej regulačné požiadavky sa vzťahujú na akékoľvek systémy AI, jej tvorcov a prevádzkovateľov, pokiaľ akýmkoľvek spôsobom zasahujú do života obyvateľov a štátnych celkov EÚ.
- **využíva pomerne širokú definíciu systému AI**, ktorá zahŕňa nielen strojové učenie, ale aj prístupy založené na logike, resp. poznaní (logic and knowledge based) a štatistické metódy (v zásade sa snaží pokryť celú problematiku symbolických a subsymbolických systémov AI).
- reguluje a zákonom **rieši vzťahy medzi prevádzkovateľmi a používateľmi** technológií AI. Ide tak o **systémy AI, ktoré nejakým spôsobom interagujú s okolím** (netýka sa to napr. uzavretého systému, ktorý riadi nejaký výrobný proces a v rámci výrobnej linky pracuje len s technickými dátami materiálov a pod.).
- podľa osvedčeného vzoru GDPR **stanovuje vysoké sankcie za porušenie nariadenia**.[448] Pri bežnom porušení sankcie siahajú až do výšky 20 miliónov Eur, resp. 4% celosvetového obratu, no ak by išlo o porušenie článku 5 (zakázané

446 General Data Protection Regulation – *Nariadenie Európskeho parlamentu a Rady (EÚ) 2016/679 z 27. apríla 2016 o ochrane fyzických osôb pri spracúvaní osobných údajov a o voľnom pohybe takýchto údajov*. [on-line]. [cit. 19. februára 2022].
Dostupné na internete: <https://eur-lex.europa.eu/legal-content/SK/TXT/?uri=celex%3A32016R0679>

447 Túto problematiku sme spomínali v kapitole 2.8.

448 Vysoké sankcie spolu s definovanými a realizovanými mechanizmami kontroly rešpektovania GDPR asi v najväčšej miere prispeli k rýchlej adaptácii potrebných postupov ochrany osobných údajov. To isté sa očakáva aj v oblasti umelej inteligencie.

systémy AI) a článku 10 (správa dát), sankcie sa môžu vyšplhať až do výšky 30 miliónov Eur, resp. 6% celosvetového obratu.

Uplatnenie pravidiel Aktu o umelej inteligencii môžeme očakávať v horizonte niekoľkých rokov, keďže návrh z apríla 2021 prechádza jeden až dvojročným legislatívnym procesom EÚ, po ktorom bude nasledovať legisvakančná lehota v dĺžke 24 mesiacov umožňujúca po schválení nariadenia jeho implementáciu. Táto legisvakančná lehota je však na rozdiel od iných regulácií dosť ošemetná, lebo systémy umelej inteligencie sú principiálne iné – z povahy veci nie je možné technologicky dopĺňať zabezpečenie požiadaviek nariadenia až po začiatku jeho účinnosti.[449]

Vzhľadom na dizajn, parametrizáciu, trénovanie a dôsledné zabezpečenie fungujúceho systému AI je treba požiadavky nariadenia implementovať od počiatku vývoja daného systému. Zverejnený návrh Aktu o umelej inteligencii tak extrémne naberá na význame, lebo – odhliadnuc od možných úprav v rámci prebiehajúceho legislatívneho procesu – bude zaväzovať už v súčasnosti vznikajúce systémy, ich tvorcov a prevádzkovateľov, resp. poskytovateľov.[450]

Akt o umelej inteligencii definuje tri typy technológií AI:

- zakázané systémy
- vysoko rizikové systémy
- ostatné systémy

Medzi zakázané systémy AI patria (čl. 5):

- podprahové techniky ovplyvňujúce správanie jednotlivcov
- metódy využívajúce slabiny zraniteľných osôb (napr. deti, hendikepovaní)
- systémy sociálneho hodnotenia (social scoring) využívané štátnou správou

449 PATTYNOVÁ, *Výzvy a právní aspekty umělé inteligence*. Dostupné na internete: <https://www.tuesday.cz/akce/umela-inteligence-1/zaznam-akce/>

450 Autor sa spolupodielal na tvorbe Smernice Katolíckej cirkvi v oblasti GDPR a následnej verifikácii, návrhu i úpravách viacerých informačných systémov Cirkvi, pri ktorých bolo či už technologickými alebo procesnými postupmi možné dosiahnuť splnenie legislatívnych požiadaviek ochrany osobných údajov. **Ak v prípade GDPR patrili princípy *security by design* a *privacy by design* k *best practices*, v prípade systémov AI je *ethics by design* a *regulation by design* nutnou podmienkou ich vývoja, nasadenia a prevádzky**. V tejto oblasti by teda systémy AI mali byť tzv. odolné voči budúcnosti (future proof).

- biometrická identifikácia v reálnom čase na verejných priestranstvách orgánmi štátnej moci (iní ani nemôžu) okrem oprávnených výnimiek (face recognition môže mať výnimku, no napr. social scoring nie)

Vysoko rizikové systémy (čl. 6 + prílohy II. a III.), na ktoré sa bude viazať systém povinností, resp. zodpovednosti a certifikácie v rámci EÚ, sa radia do viacerých tzv. sektorov:

- zamestnávanie
- vzdelávanie
- zdravotníctvo
- prístup k verejným službám
- biometrická identifikácia
- overovanie bonity osôb a pod.

Kategória Ostatné systémy AI zahŕňa všetky ostatné technológie umelej inteligencie, ktoré – až na čl. 52 – nie sú viazané žiadnymi osobitnými povinnosťami. Čl. 52 nariaďuje, aby pre používateľov bolo zrejmé, že interagujú so systémom AI. V zásade možno povedať, že Ostatné systémy AI sú viazané povinnosťou transparentnosti. Naviac sa odporúča, aby členské štáty podporovali dobrovoľné prijatie etických záväzkov, napr. prostredníctvom kódexu správania a pod.

Dôležitou súčasťou nariadenia o AI sú tzv. **významné povinnosti poskytovateľov**:[451]

- **riadenie rizík** (čl. 9): povinnosť implementovať interné procesy za účelom identifikácie, analýzy a zmierňovania následkov rizík v súvislosti s vysoko rizikovými systémami AI.
- **správa dát** (čl. 10): požaduje, aby testovacie a tréningové datasety boli „vysokej kvality“, aby tak systém AI, ktorý ich využíva, nebol diskriminačný a nevytváral nepredvídané, resp. nesprávne výsledky.
- **technická dokumentácia** (čl. 11): stanovuje povinnosť pripraviť detailnú technickú dokumentáciu (jej rozsah je definovaný v prílohe IV), ktorá umožní auditovanie systému AI, vrátane overenia jeho výsledkov.

451 PATTYNOVÁ, *Výzvy a právní aspekty umělé inteligence*.
Dostupné na internete: <https://www.tuesday.cz/akce/umela-inteligence-1/zaznam-akce/>

- **interné záznamy** (čl. 12 a 20): nariaďuje požiadavku na zaznamenávanie (logovanie) a uchovávanie záznamov jednotlivých udalostí spojených s vývojom a využívaním systému AI.
- **transparentnosť** (čl. 13): ukladá povinnosť poskytnúť používateľovi dokumentáciu a manuál pre používanie systému AI, vďaka ktorému používateľ môže porozumieť a vykonávať kontrolu nad vytváraním výsledkov systému AI.
- **ľudský dohľad** (čl. 14): vyžaduje zabezpečenie možnosti zásahu kvalifikovaných osôb určených používateľom, predovšetkým schopnosť celkom prerušiť prevádzku systému AI a zmeniť výsledok vytvorený týmto systémom.
- **presnosť, robustnosť (spoľahlivosť) a kybernetická bezpečnosť** (čl. 12): požaduje, aby miera presnosti bola deklarovaná v technickej dokumentácii, aby systém AI bol odolný voči chybám a nejasnostiam, a tiež proti škodlivým zásahom tretích strán.
- **systém riadenia kvality** (čl. 17): ukladá povinnosť implementovať systém riadenia kvality, ktorý by minimálne mal zahŕňať rozsiahlu internú dokumentáciu a procesy pre zabezpečenie testovania a preverovania.
- **monitorovanie po uvedení na trh** (post-marketing monitoring, čl. 54): vyjadruje povinnosť zabezpečiť monitorovanie v súlade s nariadením aj po uvedení do predaja, resp. prevádzky, ktoré by bolo založené na dátach poskytnutých používateľmi, resp. získaných z iných zdrojov.
- **posúdenie zhody** (čl. 43): okrem výnimiek (zdravie, bezpečnosť osôb,...) sa požaduje posúdenie vykonávané priamo poskytovateľmi systému AI (self-assessment), a tiež vykonanie jeho aktualizácie pri každej významnej zmene systému.
- **registrácia** (čl. 61): ukladá povinnosť registrovať systém AI a poskytnúť o ňom informácie podľa prílohy VII. Tieto informácie budú vedené v databáze EÚ.
- **oznamovanie incidentov** (čl. 62): stanovuje povinnosť hlásiť príslušným orgánom členských štátov EÚ všetky závažné incidenty alebo poruchy systému AI, ktoré by predstavovali porušenie povinností EÚ pre ochranu ľudských práv.

Významné povinnosti poskytovateľov podľa Aktu o umelej inteligencii prakticky vyjadrujú veľa – podľa predkladateľov nariadenia to podstatné – z návrhov a postrehov, ktoré sme

rozoberali v 2. kapitole a sumarizovali v kapitole 3.1. Z nášho pohľadu tak zapracovanie podstatných požiadaviek na eticko-právne regulácie do pripravovaného nariadenia EÚ môže byť dôvodom pre veľké uspokojenie a zadosťučinenie.

Množina významných povinností však prináša aj jedno veľké jarmo: z takmer ideálu pre riešenie dôveryhodného, transparentného a na človeka orientovaného systému AI sa týmto nariadením stáva minimálny štandard – zákonná povinnosť, ktorú nebude ľahké splniť.

Ponajprv, nie je po technickej a procesnej stránke vôbec ľahké niektoré povinnosti splniť. Uveďme malý príklad – je v súčasnosti reálne mať dokonalý tréningový dataset, alebo mať tak vytrénovaný systém AI, aby neniesol riziká *long-tail* efektu? Nájdeme viacero scenárov, ktoré poukazujú na extrémnu obťažnosť splnenia niektorých povinností...

Výzvou je i prípadný nesúlad s už existujúcim nariadením o ochrane osobných údajov (GDPR). Ako napríklad zosúladiť princíp minimalizácie údajov z GDPR s požiadavkou na úplnosť datasetov a povinnosťou uchovávania logov s citlivým obsahom? GDPR stanovuje tituly, t.j. dôvody pre spracúvanie osobných údajov. Ako obhájiť dôvody špecifické pre konkrétne algoritmy AI? Ako zosúladiť požiadavku na používateľské dáta, vďaka ktorým sa systém „učí“, s podmienkami GDPR? Viaceré problémy ešte čakajú na vyriešenie...[452]

Ďalším problémom je ekonomická, odborná i časová náročnosť spojená so zabezpečením kompatibility s uvedeným nariadením. Čím zložitejší systém a naviac ak ide o vysoko rizikový systém, tým väčšou výzvou bude implementácia požiadaviek nariadenia.

A do tretice – vyvinúť a prevádzkovať systém kompatibilný s Aktom o umelej inteligencii sa prejaví vo zvýšenej ekonomickej záťaži v porovnaní s konkurenciou. Pravda, bavíme sa o trhoch mimo EÚ, na ktoré by nemali dosah ani rozšírené teritoriálne dopady tohoto nariadenia.

Znovu sa vrátiac k analógii nariadenia o ochrane osobných údajov (GDPR) – i v prípade Aktu o umelej inteligencii sa EÚ bude potýkať s rovnakými problémami „veľkého jarma“ regulačných požiadaviek. S odstupom času možno povedať, že vzhľadom na nekontrolovateľné el-dorado netransparentného narábania s osobnými údajmi naprieč celým digitálnym svetom bolo GDPR zásahom v hodine dvanástej a jeho benefity neustále

452 PATTYNOVÁ, *Výzvy a právní aspekty umělé inteligence*.
Dostupné na internete: <https://www.tuesday.cz/akce/umela-inteligence-1/zaznam-akce/>

prichádzajú. Silný regulačný imperatív (regulácia stupňa nariadenie, vysoké pokuty a pod.) a pomerne dobre zvládnutá legisvakančná lehota naviac zabezpečili, že nariadenie prinútilo tvorcov, prevádzkovateľov a poskytovateľov informačných systémov prispôsobiť sa a akceptovať pravidlá hry.

Vzhľadom na solídny regulačný obsah Aktu o umelej inteligencii, možnosti jeho nasadenia a vymáhania podľa vzoru a na základe skúseností s GDPR si myslíme, že toto nariadenie má potenciál byť dôležitým vkladom pre zabezpečenie etického a právne rámca využívania systémov umelej inteligencie (ANI) v informačnej spoločnosti a digitálnom veku.

Akt o umelej inteligencii v súčasnosti prechádza legislatívnym konaním v rámci orgánov EÚ. Dúfajme, že jeho ovocím bude zachovanie a prípadné doplnenie dôležitých regulačných princípov, okorenené schopnosťou nájsť riešenia tých parciálnych problémov, ktoré by inak celoplošné nasadenie nariadenia diskvalifikovali.

I keď má Akt o umelej inteligencii potenciál širokospektrálneho pokrytia využívania technológií umelej inteligencie v modernej spoločnosti, už z princípu nemôže pokryť dve dôležité oblasti: nasadenie v oblasti pokročilého riadenia štátu, spravodajstva a dohľadu a nasadenie vo vojenskej oblasti.[453]

Vzhľadom na špecifikum a dosah nasadenia systémov AI v spravodajských službách[454] pre ľudské práva, ochranu demokracie a slobôd si myslíme, že táto oblasť by mala byť pokrytá už základnými legislatívnymi mechanizmami a verejným dohľadom demokratickej spoločnosti, ktoré sa týkajú pôsobenia spravodajských služieb vo všeobecnosti.

Keďže akékoľvek obmedzovanie technológií umelej inteligencie vo vojenskej oblasti môže byť chápané ako bezpečnostné riziko a zníženie bojaschopnosti modernej armády, jednostranne prijaté regulácie nemusia byť účinné – nielen pre to, že sa jednou stranou ťažko prijímajú (aj keď pre hodnotovo orientovanú spoločnosť by to malo byť povinnosťou), ale i pre malú šancu na ich extra teritoriálne rozšírenie a akceptovanie. To však neznamená, že by sme mali na eticko-právne regulácie AI v armáde rezignovať. Veď v predchádzajúcich kapitolách spomenuté eticko-právne procesy prebiehajúce napríklad v armáde USA môžu byť základom pre celosvetovú diskusiu mocností a na základe tlaku

453 Čl. 2, 3: „Toto nariadenie sa neuplatňuje na systémy umelej inteligencie vyvinuté alebo používané výlučne na vojenské účely.“

454 Aspekty nasadenia systémov AI v oblasti spravodajstva a algokracie sme diskutovali v kapitole 2.6.

verejnosti, angažovanosti jednotlivých častí spoločnosti v rôznych regiónoch sveta i úsilia zodpovedných strán môžu viesť k prijatiu celosvetových pravidiel i záväzkov pre oblasť vývoja a nasadenia rizikových vojenských systémov vybavených technológiami umelej inteligencie. Vývoj a nasadenie viacerých extrémne nebezpečných vojenských technológií je v súčasnosti na základe vzájomného konsenzu a právnych záväzkov zakázané, je preto žiadúce, aby sa tak stalo aj v prípade niektorých scenárov nasadenia smrtiacich automatických zbraňových systémov a ďalších rizikových technológií využívajúcich algoritmy umelej inteligencie.[455]

V akčnom rádiuse doteraz uvedených legislatívnych aktivít sa prevažne nachádzajú slabé systémy umelej inteligencie (ANI). **I keď základný eticko-právny rámec zostáva v platnosti, nie je jednoduché právne uchopiť systémy, ktoré sa aspoň v niektorých aspektoch približujú silnej umelej inteligencii (AGI). Problémy pramenia i z technologickej náročnosti daného predmetu práva[456] i z celého komplexu otvorených otázok kognitívnych schopností a modelov správania, aj keď sa v súčasnosti ešte nejedná o právne reakcie na teóriu mysle alebo simulácie emócií a vedomia AGI**.

V prípade technológií silnej a všeobecnej umelej inteligencie bude vyvstávať otázka ich právneho postavenia.[457] Mohla by za istých okolností existovať určitá spôsobilosť systému AGI byť nositeľom práv a povinností predvídaných právnym poriadkom? Prípadne, mohol by systém AGI vlastnými (právnymi?) úkonmi nadobúdať práva a brať na seba povinnosti? Ako by to bolo v prípade schopností analogických ľudskému uvažovaniu a prejavovaniu vôle? Kto by bol nositeľom autorských práv v prípade diel

455 Tieto scenáre sme rozoberali v kapitole 2.7.

456 Autor sa aj u renomovaných právnických agentúr stretol s analytickými a aplikačnými medzerami, ktoré vyplývali z nedostatku pochopenia princípov a fungovania zložitejších systémov AI.

457 Už v roku 2017 Európsky parlament, riešiac problematiku robotických systémov, reflektoval problematiku právneho statusu AI vo výzve Komisii, aby sa venovala vytvoreniu „špecifického právneho postavenia pre roboty v dlhodobom horizonte, aby sa aspoň tie najsofistikovanejšie autonómne roboty mohli považovať za elektronické osoby zodpovedné za náhradu akejkoľvek škody, ktorú môžu spôsobiť.“ *Uznesenie Európskeho parlamentu zo dňa 16. 2. 2017 obsahujúce odporúčania pre Komisiu k normám občianskeho práva v oblasti robotiky (2015/2103(INL))*. [on-line]. [cit. 29. marca 2022]. Dostupné na internete: <https://eur-lex.europa.eu/legal-content/SK/TXT/PDF/?uri=CELEX:52017IP0051&from=EN>

vytvorených algoritmami umelej inteligencie?[458]

I tieto otázky poukazujú na komplexnosť výzvy a potencionálny dopad budúcich systémov AGI prakticky na akúkoľvek oblasť života spoločnosti.

3.4. Aktivity Cirkvi

Naliehavosť potreby skúmať, navrhnúť, prijať a realizovať etický rámec vývoja, používania a fungovania technológií umelej inteligencie rezonuje aj v Katolíckej cirkvi. Mnohorakým spôsobom je táto téma uchopená v rámci akademického prostredia, predovšetkým v interdisciplinárnom poňatí.[459] V agende vatikánskych inštitúcií je problematika AI pokračovaním angažovanosti v sociálnej sfére, nadväzujúc napríklad na širšie koncipované aktivity Pápežskej akadémie pre život, ktoré sme spomínali v úvode tejto publikácie.[460]

Doteraz najširšie koncipovaným počinom Pápežskej akadémie pre život[461] v oblasti etiky umelej inteligencie bola vo februári 2020 zorganizovaná konferencia **renAIssance 2020**.[462]

458 ŠTARHA, Š., GAŠPAROVIČ, R. *AI z pohľadu práva*. [on-line]. [cit. 29. marca 2022]. Dostupné na internete: <https://www.epravo.sk/top/clanky/ai-z-pohladu-prava-4483.html>

459 Jedným z príkladom môže byť napríklad nedávny webinár "Man, Machine, and the Future of AI", ktorý v rámci série prednášok Conversations That Matter: The Crossroads of Science and Human Dignity Fall 2021 organizoval McGrath Institute for Church Life, University of Notre Dame.
Conversations That Matter: The Crossroads of Science and Human Dignity Fall 2021. [on-line]. [cit. 28. marca 2022].
Dostupné na internete: <https://mcgrath.nd.edu/conferences/academic-pastoral/conversations-that-matter-the-crossroads-of-science-and-human-dignity/conversations-that-matter-the-crossroads-of-science-and-human-dignity-fall-2021/>

460 Išlo o podujatie Hackaton 2018, ktoré sa venovalo jednému z vážnych sociálnych problémov informačnej spoločnosti a digitálneho veku – digitálnemu rozdeleniu (digital divide). V duchu hesla „zapojenie mladých a technológií v prospech spoločného dobra“ sa v kontexte využitia technológií hľadali odpovede na otázky sociálneho začlenenia, medzi náboženského dialógu a zdrojov pre utečencov.
Vatican Hackathon – harnessing youth, technology to serve common good. [on-line]. [cit. 28. marca 2022].
Dostupné na internete: <https://www.vaticannews.va/en/vatican-city/news/2018-03/vatican-hackathon--.html>

461 Dostupné na internete: <https://www.academyforlife.va/>

462 Konferencia bola známa aj pod názvom Rome Call for AI Ethics, či na sociálnych sieťach pod značkou #RomeCallforAIethics.

Prestížny AI Index Report publikovaný každoročne inštitútom HAI of Stanford University zaradil túto konferenciu medzi najhorúcejšie témy roku 2020.[463]

Na konferencii, ktorá sa konala 26. - 28. februára 2020 v Ríme, vystúpilo viacero špičkových odborníkov na etiku umelej inteligencie z akademickej oblasti i vedúcich predstaviteľov veľkých korporácií z oblasti informačných technológií.[464]

Jedným z výsledkov konferencie bolo aj podpísanie **Výzvy na etiku v umelej inteligencii**[465] so zámerom „**akcentovať etický prístup k umelej inteligencii a podporovať zmysel pre zodpovednosť medzi organizáciami, vládami a inštitúciami s cieľom vytvoriť budúcnosť, v ktorej digitálne inovácie a technologický pokrok slúžia ľudským schopnostiam a tvorivosti, a nie ich postupnému nahrádzaniu**“.[466]

Signatármi výzvy boli Vincenzo Paglia, predseda Pápežskej akadémie pre život, Brad Smith, prezident spoločnosti Microsoft, John Kelly III, viceprezident IBM, Qu Dongyu, generálny riaditeľ FAO a Paola Pisano, ministerka inovácií Talianska.

Zámerom organizátora, Pápežskej akadémie pre život, bol **vstup do diskusie o etických kritériách AI v jednotlivých oblastiach vývoja a nasadenia**. K nosným témam patrila **diskusia zameraná na snahu o premostenie pohľadov jednotlivých krajín, firiem, záujmových skupín na oblasť etiky umelej inteligencie**, pričom naplneným cieľom konferencie sa stalo **stanovenie konkrétnych etických kritérií**.

Sumarizáciu tohoto úsilia vyjadruje vyššie spomenutý záverečný dokument, ktorého súčasťou je identifikovanie troch oblastí vplyvu a formulovanie šiestich princípov pre použitie umelej inteligencie pre dobro človeka a bez strachu zo zneužitia.

Akcentované oblasti vplyvu sú vo výzve definované nasledovne:

---

463 *Measuring trends in Artificial Intelligence – 2021 AI Index Report*. [on-line]. [cit. 28. marca 2022]. Dostupné na internete: <https://aiindex.stanford.edu/ai-index-report-2021/>

464 *The „good“ algorithm*. [on-line]. [cit. 28. marca 2022]. Dostupné na internete: <https://www.academyforlife.va/content/dam/pav/documenti%20pdf/2020/Assemblea/Atti_Assemblea_e_28febbraio/Atti%20completi_PAV_2020_.pdf>

465 *Rome Call for AI Ethics (document)*. [on-line]. [cit. 28. marca 2022]. Dostupné na internete: <https://www.romecall.org/wp-content/uploads/2022/03/RomeCall_Paper_web.pdf>

466 *Rome Call for AI Ethics*. [on-line]. [cit. 19. augusta 2020]. Dostupné na internete: <http://www.academyforlife.va/content/pav/en/events/intelligenza-artificiale.html>

- **etika** – všetky ľudské bytosti sa rodia slobodné a rovné v dôstojnosti a právach.
- **vzdelávanie** – premeniť svet prostredníctvom inovácie umelej inteligencie znamená zaviazať sa budovať budúcnosť pre mladé generácie a s nimi.
- **práva** – vývoj technológií AI v službách ľudstva a planéty sa musí odrážať v predpisoch a zásadách, ktoré chránia ľudí, najmä slabých a znevýhodnených, a prírodné prostredie.

Považujeme za vhodné uvedené oblasti vplyvu popísať priamo slovami záverečného dokumentu *Rome Call for AI Ethics*, nakoľko ide o hutný text sumarizujúci postrehy, návrhy a závery takmer tristo stranového konferenčného zborníka.

Oblasť etiky stavia na všeobecnej deklarácii OSN o ľudských právach. Všetky ľudské bytosti sa rodia slobodné a rovné v dôstojnosti a právach. Sú obdarené rozumom a svedomím a mali by voči sebe konať v duchu spolupatričnosti (por. čl. 1 Dekrétu OSN o ľudských právach). Táto základná podmienka slobody a dôstojnosti musí byť chránená a zaručená aj pri výrobe a používaní systémov umelej inteligencie. To sa musí uskutočniť zabezpečením práv a slobody jednotlivcov tak, aby neboli algoritmami diskriminovaní z dôvodu svojej „rasy, farby pleti, pohlavia, jazyka, náboženstva, politického alebo iného zmýšľania, národného alebo sociálneho pôvodu, majetku, rodu alebo iného postavenia“ (čl. 2).[467]

Systémy AI musia byť koncipované, navrhnuté a implementované tak, aby slúžili a chránili ľudské bytosti a prostredie, v ktorom žijú. Tento základný pohľad sa musí premietnuť do záväzku vytvoriť také životné podmienky (spoločenské aj osobné), ktoré umožnia skupinám aj jednotlivým členom snažiť sa o plné vyjadrenie, ak je to možné.[468]

Aby bol technologický pokrok v súlade so skutočným pokrokom ľudského rodu a úctou k planéte, musí spĺňať tri požiadavky. Musí zahŕňať každú ľudskú bytosť a nikoho nediskriminovať; musí mať na zreteli dobro ľudstva a dobro každej ľudskej bytosti; napokon musí brať do úvahy zložitú realitu nášho ekosystému a vyznačovať sa tým, ako sa stará o planétu (náš „spoločný a zdieľaný domov“) a chráni ju vysoko udržateľným spôsobom, ktorý zahŕňa aj využívanie umelej inteligencie pri zabezpečovaní udržateľných

467 *Rome Call for AI Ethics (document)*. [on-line]. Str. 5. [cit. 28. marca 2022]. Dostupné na internete: <https://www.romecall.org/wp-content/uploads/2022/03/RomeCall_Paper_web.pdf>

468 *Rome Call for AI Ethics (document),* s. 5.

potravinových systémov v budúcnosti. Okrem toho si každý človek musí byť vedomý toho, kedy komunikuje so strojom.[469]

Technológia založená na umelej inteligencii sa nikdy nesmie používať na akékoľvek zneužívanie ľudí, najmä tých najzraniteľnejších. Namiesto toho sa musí používať na pomoc ľuďom pri rozvíjaní ich schopností a na podporu planéty.[470]

Oblasť vzdelávania je pozvaním premieňať svet prostredníctvom inovácie umelej inteligencie, a tak sa zaviazať budovať budúcnosť pre mladé generácie a s nimi. Tento zámer sa musí odraziť v záväzku k vzdelávaniu, vo vypracovaní špecifických učebných osnov, ktoré zahŕňajú rôzne humanitné, vedecké a technické disciplíny, a v prevzatí zodpovednosti za vzdelávanie mladších generácií. Tento zámer tiež znamená snahu o zlepšenie kvality vzdelávania, ktorého sa mladým ľuďom dostáva; toto vzdelávanie sa musí uskutočňovať prostredníctvom metód, ktoré sú prístupné pre všetkých, ktoré nediskriminujú a ktoré môžu ponúknuť rovnosť príležitostí a zaobchádzania. Všeobecný prístup k vzdelávaniu sa musí dosiahnuť prostredníctvom zásad solidarity a spravodlivosti.[471]

Prístup k celoživotnému vzdelávaniu musí byť zaručený aj pre starších ľudí, ktorým sa musí ponúknuť možnosť prístupu k off-line službám počas digitálneho a technologického prechodu. Okrem toho sa tieto technológie môžu ukázať ako nesmierne užitočné pri pomoci ľuďom so zdravotným postihnutím, aby sa mohli učiť a stať sa nezávislejšími: inkluzívne vzdelávanie preto znamená aj využívanie umelej inteligencie na podporu a integráciu každého človeka, poskytovanie pomoci a možností sociálnej participácie (napr. práca na diaľku pre osoby s obmedzenou mobilitou, technologická podpora pre osoby s kognitívnymi poruchami atď.).[472]

Vplyv transformácií, ktoré priniesol fenomén umelej inteligencie v spoločnosti, v práci a vo vzdelávaní, si vynútil prepracovanie školských osnov, aby sa motto vzdelávania "nikto nezostane pozadu" stalo skutočnosťou. V oblasti vzdelávania sú potrebné reformy s cieľom zaviesť vysoké a objektívne normy, ktoré môžu zlepšiť individuálne výsledky. Tieto normy by sa nemali obmedzovať na rozvoj digitálnych zručností, ale mali by sa

469 *Rome Call for AI Ethics (document),* s. 6.

470 *Rome Call for AI Ethics (document),* s. 6.

471 *Rome Call for AI Ethics (document),* s. 7.

472 *Rome Call for AI Ethics (document),* s. 7.

zamerať na to, aby každý človek mohol naplno prejaviť svoje schopnosti a pracovať pre dobro komunity, aj keď z toho nemá osobný prospech.[473]

Pri navrhovaní a plánovaní spoločnosti zajtrajška musí využívanie systémov AI sledovať formy činnosti, ktoré sú sociálne orientované, tvorivé, prepojené, produktívne, zodpovedné a schopné pozitívne ovplyvniť osobný a spoločenský život mladších generácií. Sociálny a etický rozmer umelej inteligencie musí byť aj jadrom vzdelávacích aktivít v oblasti AI.[474]

Hlavným cieľom tohto vzdelávania musí byť zvyšovanie povedomia o príležitostiach a tiež o možných kritických problémoch, ktoré fenomén AI predstavuje z hľadiska sociálneho začlenenia a rešpektu k individuálnej osobe.[475]

Oblasť práv sumarizuje, ako sa rozvoj umelej inteligencie v službách ľudstva a planéty musí odrážať v predpisoch a zásadách, ktoré chránia ľudí (najmä slabých a znevýhodnených) a prírodné prostredie. Etický záväzok všetkých zúčastnených strán je kľúčovým východiskom na to, aby sa táto budúcnosť stala skutočnosťou, pričom sú absolútne nevyhnutné hodnoty, zásady a v niektorých prípadoch aj právne predpisy, ktoré tento proces podporujú, systematizujú a usmerňujú.[476]

Na vývoj a implementáciu systémov umelej inteligencie, ktoré budú prospešné pre ľudstvo a planétu a zároveň budú slúžiť ako nástroje na budovanie a udržiavanie medzinárodného mieru, musí ísť vývoj umelej inteligencie ruka v ruke so spoľahlivými opatreniami v oblasti digitálnej bezpečnosti.[477]

Aby technológie AI mohli fungovať ako nástroj v prospech ľudstva a planéty, musíme do centra verejnej diskusie postaviť tému ochrany ľudských práv v digitálnej ére. Nastal čas položiť si otázku, či nové formy automatizácie a algoritmickej činnosti nevyžadujú vytvorenie silnejších zodpovedností. Predovšetkým bude nevyhnutné zvážiť určitú formu „povinnosti vysvetľovania“: musíme sa zamyslieť nad tým, aby boli zrozumiteľné nielen kritériá rozhodovania algoritmických agentov založených na umelej inteligencii, ale aj ich účel a ciele. Tieto zariadenia musia byť schopné ponúknuť jednotlivcom informácie o logike algoritmov používaných na rozhodovanie. Tým sa zvýši nielen transparentnosť,

473 *Rome Call for AI Ethics (document),* s. 7-8.

474 *Rome Call for AI Ethics (document),* s. 8.

475 *Rome Call for AI Ethics (document),* s. 8.

476 *Rome Call for AI Ethics (document),* s. 9.

477 *Rome Call for AI Ethics (document),* s. 9.

sledovateľnosť a zodpovednosť, ale i adekvátnosť a váha rozhodovacieho procesu podporovaného počítačom. Je potrebné podporovať nové formy regulácie na podporu transparentnosti a dodržiavania etických zásad, najmä v prípade pokročilých technológií, ktoré majú vyššie riziko vplyvu na ľudské práva, ako je napríklad rozpoznávanie tváre.[478]

Praktickým vyjadrením konferenčných záverov zhrnutých v uvedených troch oblastiach sa stalo šesť princípov využitia systémov AI pre dobro človeka a bez strachu zo zneužitia:

- **transparentnosť** – algoritmy AI musia byť transparentné a zrozumiteľné pre všetkých.
- **inklúzia** – je potrebné zohľadniť potreby všetkých ľudí, aby z toho mali prospech všetci a aby sa všetkým jednotlivcom poskytli čo najlepšie podmienky na sebarealizáciu a rozvoj. Systémy AI nesmú nikoho diskriminovať, pretože každý človek má rovnakú dôstojnosť.
- **zodpovednosť** – tí, ktorí navrhujú a zavádzajú používanie technológií AI, musia postupovať zodpovedne a transparentne. Vždy musí existovať niekto, kto preberá zodpovednosť za to, čo systém AI vykonáva.
- **nestrannosť** – netvoriť a nekonať na základe zaujatosti, a tak chrániť spravodlivosť a ľudskú dôstojnosť. Systémy AI sa predsudkami nesmú nechať viesť alebo ich vytvárať.
- **spoľahlivosť** – systémy AI musia byť schopné spoľahlivo fungovať.
- **bezpečnosť a súkromie** – technológie AI musia byť bezpečné a rešpektovať súkromie používateľov.

Cieľ, ktorý konferenciou *renAIssance* Pápežská akadémia pre život sledovala, vyjadril Mons. Vincenzo Paglia slovami: „Zámerom výzvy je vytvoriť hnutie, ktoré sa rozšíri a zapojí ďalších aktérov: verejné inštitúcie, mimovládne organizácie, priemyselné odvetvia a skupiny, aby určili smer vývoja a používania technológií odvodených od umelej inteligencie. Z tohto hľadiska môžeme povedať, že prvý podpis tejto výzvy nie je vyvrcholením, ale východiskom pre záväzok, ktorý sa javí ako ešte naliehavejší a dôležitejší než kedykoľvek predtým. Pripojenie sa k tejto iniciatíve pre zástupcov priemyslu, ktorí ju podpíšu, znamená záväzok, ktorý má význam aj z hľadiska nákladov a priemyselného príspevku k vývoju a distribúcii ich výrobkov. Ak sa Akadémia cíti byť

478 *Rome Call for AI Ethics (document),* s. 9.

povolaná zintenzívniť svoje úsilie o uľahčenie získavania poznatkov a podpisov iných medzinárodných aktérov, táto výzva je len prvým krokom, po ktorom by mali nasledovať ďalšie. Text výzvy sa vyznačuje aj tým, že je prvým **pokusom sformulovať súbor etických kritérií so spoločnými referenčnými bodmi a hodnotami**, čím ponúka príspevok k rozvoju spoločného jazyka na interpretáciu toho, čo je človek“.

Celá výzva je v duchu kresťanskej antropológie intenzívne zameraná na človeka. Ide o viackrát spomínaný *human-centered* prístup, ktorý je však v záveroch konferencie širšie koncipovaný s osobitným zreteľom na sociálnu spravodlivosť, rovnosť a ľudskú dôstojnosť.

Dôležitým vyjadrením je **požiadavka, aby etický záväzok bol základom pre regulácie a právne normy**.

V niektorých aspektoch badať jasný vplyv osobitne v poslednej dekáde prebiehajúcich etických a legislatívnych procesov, ktoré boli integrované do Výzvy na etiku v umelej inteligencii a zároveň sa stali aj súčasťou návrhu európskeho Aktu o umelej inteligencii. Keďže finálna verzia návrhu Aktu o umelej inteligencii bola predstavená v apríli 2021, bolo by zaujímavé sledovať, ako a do akej miery obsah rímskej Výzvy na etiku v AI ovplyvnil tento regulačný návrh EÚ.

V súčasnosti najucelenejší a v minulej kapitole predstavený regulačný rámec pre technológie umelej inteligencie, Akt o umelej inteligencii, tak stavia na etických a hodnotových základoch jasne korešpondujúcich s rímskou výzvou, pričom každý z nich – Akt i Výzva – má svoje vlastné zameranie. Akt ako nariadenie EÚ je striktným právnym rámcom vyžadujúcim dodržiavanie etických a právnych noriem. Výzva – ako uvádza Mons. Paglia – si kladie za cieľ iniciovať univerzálne hnutie, naprieč celým svetom i všetkými oblasťami spoločnosti, ktoré chce usmerňovať a formovať vývoj a používanie technológií umelej inteligencie podľa etických princípov a morálnych hodnôt pre dobro človeka, spoločnosti a sveta, akcentujúc dôstojnosť každého človeka a ochranu životného prostredia ako spoločného domu celého ľudstva.

Význam snahy o celosvetové hnutie, na ktorom aktívne participuje Katolícka cirkev je o to väčší, ak si uvedomíme hodnotovú a etickú diskrepanciu vo svete. Podľa štúdie *The global landscape of AI ethics guidelines* **v zásade panuje celospoločenský konsenzus ohľadom potreby etických usmernení v oblasti umelej inteligencie, ktorý je však sprevádzaný podstatnými rozdielmi v chápaní etiky a hodnôt**: „Naše výsledky ukazujú globálnu konvergenciu, ktorá sa objavuje okolo piatich etických princípov (transparentnosť,

spravodlivosť a férovosť, neškodnosť, zodpovednosť a súkromie), pričom sa podstatne líšia v tom, ako sa tieto zásady interpretujú; prečo sa považujú za dôležité, akej problematiky, oblasti alebo aktérov sa týkajú a ako sa by sa mali uplatňovať. Naše zistenia zdôrazňujú význam integrácie usmernení s podstatnou etickou analýzou a primeranou implementáciou stratégií.“[479]

I keď pre skutočne etický vývoj a nasadenie systémov AI sú podstatné tak komplexné i dosah majúce legislatívne rámce, ako je nariadenie EÚ Akt o umelej inteligencii, konferencia Pápežskej akadémie pre život, jej záber a výsledok nám správne pripomínajú celostvetové aktivity, angažovanosť a hnutie, ktoré nielenže vždy stojí na počiatku hľadania adekvátnych etických právnych noriem, ale je i katalyzátorom ich návrhu a prijatia, nehovoriac o raste celospoločenského povedomia.

Rímska konferencia renAIssance 2020 len potvrdila, že Katolícka cirkev z hĺbky svojej náuky má čím prispieť do tohto spoločného diela rozvoja moderného sveta.

I v tomto prípade treba však uviesť, že spomínané aktivity a pomerne jasné i komplexné riešenia sa primárne týkajú systémov slabej umelej inteligencie (ANI). Riešenie výziev, ktoré sa viažu na silnú, resp. všeobecnú umelú inteligenciu (AGI), je v súčasnosti doménou skôr akademického výskumu a interdisciplinárneho bádania s prienikom do oblasti transhumanizmu, neurobioetiky,[480] kognitívnych vied a pod.

Sumarizujúc tému umelej inteligencie a etiky si uvedomujeme viaceré výzvy v tejto oblasti:

- ako sa principiálne k fenoménu umelej inteligencie postavíme
- ako ju uchopíme, aký rámec zvolíme
- aké všeobecné a základné etické princípy stanovíme
- aké parciálne usmernenia či špecifické odporúčania vyjadríme
- aké možnosti pre angažovanie sa Cirkvi identifikujeme

479 JOBIN, A., IENCA, M., VAYENA, E. *The global landscape of AI ethics guidelines*. In: *Nat Mach Intell*. [on-line]. 2019, 1, s. 389–399. [cit. 28. marca 2022]. Dostupné na internete: <https://doi.org/10.1038/s42256-019-0088-2>

480 Týmto otázkam sa napríklad venujú na Oddelení neurobioetiky, Ateneo Pontificio Regina Apostolorum. Dostupné na internete: <https://www.upra.org/en/ricerca/gruppi-di-ricerca/neurobioetica/>

4. Navrhnuté riešenie etických problémov ANI

Nič veľkého nevstúpi do života smrteľníkov bez prekliatia.[481]

Začínať kapitolu týmito slovami dramatika, vzhľadom na svoju tvorbu priam tragéda, naviac politika i stratéga v jednej osobe, Sofoklesa je skutočne deprimujúce a možno trochu odzrkadľujúce naše zameranie sa na problémy a temnú stránku súčasných technológií umelej inteligencie.

Kladieme si však otázku, či to tak skutočne musí byť? Ide len o vyjadrenie dejinnej skúsenosti s ľudskou slabosťou neschopnou odolať pokušeniu, ktoré je tým väčšie, čím väčší je potenciál novej veci, ktorú spoločnosť objavila? Alebo ide skôr o sarkazmus, z čias antiky na svetlo sveta vytiahnutý, vyjadrujúci spôsob, akým sa moderný človek zmieruje s realitou, ktorá sa nedá zmeniť?

Skúsme to teda inak...

„A Boh videl všetko, čo urobil, a hľa, bolo to veľmi dobré.“

4.1. Základný postoj vo svetle Zjavenia[482]

Základné posolstvo o stvorení sveta a človeka komplexne hovorí o stvorení, že „to bolo dobré“.[483] Zároveň človek bol stvorený na Boží obraz (צֶלֶם אֱלֹהִים) s rozumom a slobodnou vôľou, aby spravoval zverený svet.[484] Človek ako obraz Boha bol uschopnený komunikovať so svojím Stvoriteľom i s ďalšími ľuďmi. K prirodzenému zákonu, ktorý je vpísaný do ľudskej duše a svedomia dostáva i pozitívny Boží zákon – Desatoro, ktoré

481 Sofokles.

482 Kapitola 4.1. je z veľkej časti prevzatá z licenciátskej práce autora, konkrétne z kapitoly 3.1. *Principiálny pohľad na základe Božieho zjavenia*.
ŠANTAVÝ, *Analýza virtuálneho sveta v kontexte náboženskej formácie, pôsobenia a života Cirkvi,* [on-line], s. 46-48. [cit. 14. marca 2022].
Dostupné na internete:
<https://peter.santavy.cloud/data/uploads/docs/analyza_virtualneho_sveta_v_kontexte_nabozenskej_formacie_posobenia_a_zivota_cirkvi.pdf>

483 „A Boh videl všetko, čo urobil, a hľa, bolo to veľmi dobré.“ (Gn 1,31).

484 „Ploďte a množte sa a naplňte zem! Podmaňte si ju a panujte nad rybami mora, nad vtáctvom neba a nad všetkou zverou, čo sa hýbe na zemi!“ (Gn 1,28).

opisuje človeka, nielen stvoreného na obraz Boží, ale človeka, ktorý tento obraz i žije: **vedome sa rozhoduje a koná tak, aby budoval vzťah voči svojmu Stvoriteľovi a múdro užíval zverený svet**.

Užívanie (podmanenie) zvereného sveta je v Gn 1,28 vyjadrené hebrejskými výrazmi „podmaňte si ju“ (וְכִבְשֻׁהָ) a „panujte“ (וּרְדוּ), ktoré sú odvodené od starostlivosti hlavy rodiny, či pastierskeho „užívania“ stáda, t.j. od **schopnosti užívať, rozvíjať a z generácie na generáciu si odovzdávať stvorený svet**.[485]

Tento základný pohľad hovoriaci o principiálnom postoji k dielu stvorenia, spravovaní a rozvoji tohoto sveta sa nesie naprieč celým starozákonným zjavením a vrcholí *v ketubím* – v múdroslovnej literatúre Starého zákona, konfrontujúcej *chokmá*, t.j. Božiu múdrosť so *sofia*, t.j. múdrosťou vtedajšieho moderného helénskeho sveta.[486] **Ovocím tejto konfrontácie nie je zavrhnutie civilizačného rozvoja, ale neustále pozvanie rozvíjať a užívať svet podľa Božích zákonov, vo vzťahu voči Bohu a pre dobro ľudí**.

Výzvy Ježiša Krista „Choďte teda, učte všetky národy...“ (Mt 28, 19) a „Choďte do celého sveta a hlásajte evanjelium všetkému stvoreniu“ (Mk 16, 15) vyjadrujú **nadčasové poslanie, zahŕňajúce civilizačný rozvoj**, v ktorom „Budete mi svedkami v Jeruzaleme i v celej Judei aj v Samárii a až po samý kraj zeme“ (Sk 1,8).

Novozákonný pohľad, spojený s ohlasovaním Evanjelia a prvopočiatkami pastoračnej starostlivosti, je veľmi pekne reprezentovaný konaním sv. Pavla v aténskom Areopágu, ktorý bol miestom koncentrácie vtedajšieho mozgového trustu – pôvodne miesto najvyššieho súdu, no v tom čase i miesto mysliteľov, predstaviteľov tej najlepšej filozofickej tradície helénskej kultúry a zároveň zástancov celého vtedajšieho Panteónu.[487] Išlo o miesto, ktoré sa pre ohlasovateľa evanjelia Pavlových kvalít a apoštola národov stalo nemalou výzvou zvestovať Ježiša ako Krista a Pána.

V sedemnástej kapitole Skutkov apoštolov môžeme veľmi jasne vidieť, ako si Pavol počínal: využil svoje vzdelanie, znalosť gréckej filozofie a kultúry, aby tak spôsobom vlastným zhromaždeným Aténčanom začal svoje zvestovanie. Vieme, ako to prebiehalo – Pavlovo priblíženie sa ich zmýšľaniu malo úspech: dali mu slovo a on mohol začať svoju

485 Podobne i v Gn 2,15 Boh, keď umiestňuje človeka v raji, dáva mu poslanie „aby ho obrábal (לְעָבְדָהּ) a strážil (לְשָׁמְרָהּ)“.

486 Kniha Múdrosti a Kniha Sirachovcova.

487 Pavol vystúpil v aténskom Areopágu počas svojej Druhej misijnej cesty (r. 50 – 52).

evanjelizačnú reč. Kristovo zmŕtvychvstanie však bolo pre ľudí odchovaných mliekom platónskej filozofie prisilnou kávou a Pavol nakoniec končí fiaskom: „vypočujeme ťa o tom inokedy“ (Sk 17,32). I keď Písmo spomína, že niektorí predsa len uverili na jeho kázanie, bola to slabá náplasť na Pavlovo sklamanie, ktoré je krásne čitateľné v niektorých chvíľach ďalšieho Pavlovho misijného putovania (napr. 1Kor 2,1-5).

I keď mohol sv. Pavol predpokladať, že jeho hlásanie v Areopágu asi skončí fiaskom, predsa si uvedomoval, že s Evanjeliom nemôže obísť intelektuálne a kultúrne stredisko vtedajšieho sveta, ktoré sa podieľalo na modernom rozvoji starovekého stredomoria a helénskej kultúry. Vedel, že treba i naďalej v tomto úsilí pokračovať,[488] a že po ňom prídu ďalší, ktorí budú prinášať Kristovo svetlo do odkrytých oblastí poznania[489] a konfrontovať i spájať moderné bádanie a rozvoj spoločnosti s Božím zjavením.[490]

Práve na základe udalostí v Areopágu v kontexte ohlasovania Evanjelia sv. Ján Pavol II. vo svojej encyklike Redemptoris missio o stálej platnosti misijného poslania poukazuje na moderné areopágy, ktoré sú priamou výzvou na permanentne platné pozvanie k misii ad gentes. Veľmi jasne definuje svet komunikácie ako prvý areopág súčasnosti[491], ktorý sa podstatne podieľa na vytváraní global village – celosvetovej dediny, v rámci ktorej moderné komunikačné prostriedky presahujú rámec nástroja komunikácie a podieľajú sa na vytváraní nového kultúrneho kontextu – novej kultúry, techniky, vyjadrovania a psychológie vzťahov – jednoducho povedané na vytváraní niečoho, čo je predzvesťou informačnej spoločnosti. Tej spoločnosti, ktorej paradigmatická zmena zahŕňa aj čo najširšiu implementáciu technológií umelej inteligencie.[492]

I keď od uvedenia encykliky už ubehlo skoro tridsať rokov, v pohľade na rodiacu sa informačnú spoločnosť s jej technologickým portfóliom môžeme vidieť nadčasový záber a prorocké slová pápeža v dobe, keď pre mnohých vtedajších mysliteľov a vizionárov

488 Pavol počas Tretej misijnej cesty v Efeze vyučoval v škole gréckeho filozofa a učiteľa rétoriky Tyranna. Por. (Sk 19,9).

489 Sv. filozof a mučeník Justín, predstavitelia alexandrijskej a antiochiskej školy, scholastici, ...

490 Skvelým zdrojom informácií o civilizačnom rozmere pôsobenia Cirkvi je: WOODS, E. T. Jr. *Ako Katolícka cirkev budovala západnú civilizáciu*. Bratislava: Redemptoristi - Slovo medzi nami, 2010.

491 JÁN PAVOL II. *Redemptoris Missio*. Praha: Zvon, 1994, s. 45-46.

492 Umelú inteligenciu v kontexte paradigmatickej zmeny informačnej spoločnosti sme uvádzali v kapitole 1.11.

virtuálny svet a kybernetický priestor, informačné a komunikačné prostriedky i algoritmy umelej inteligencie boli ťažko uchopiteľnými fenoménmi v kontexte rozvoja spoločnosti.

Analyzujúc súčasný svet, v ktorom sa technológie umelej inteligencie stávajú neodmysliteľnou súčasťou informačnej spoločnosti, môžeme s plnou zodpovednosťou označiť aj oblasť umelej inteligencie areopágom, fórom moderného veku, do ktorého – nasledujúc príklad svätého apoštola Pavla – treba s odvahou vstúpiť a vo svetle evanjelia i v tejto oblasti prispievať k nekončiacemu úsiliu budovať spravodlivý svet, chrániť dôstojnosť každého človeka a rozvíjať civilizáciu lásky.

Inak povedané, aj technológie umelej inteligencie sú súčasťou nášho dejinného kontextu, v ktorom sme pozvaní mať účasť na budovaní Božieho kráľovstva.

4.2. Interdisciplinárny rámec ako základ

Máme za to, že **skutočné riešenie etických problémov a výziev technológií umelej inteligencie nie je možné úspešne realizovať bez interdisciplinárneho rámca, v rámci ktorého sme dostatočne oboznámení aj s technologickou stránkou týchto systémov a psychologickými, sociologickými i právnymi aspektmi ich nasadenia**.

Dostatočné oboznámenie sa s technologickou stránkou nám predovšetkým umožňuje pochopiť podstatu a rozsah technologických limitov a rizík algoritmov AI, a následne si adekvátnejšie predstaviť ich dosah na jednotlivé oblasti reálneho nasadenia, dôsledky na život človeka a fungovanie spoločnosti.

Naviac sa domnievame, že pre kvalifikovanú filozofickú a teologickú diskusiu o podstatných aspektoch umelej inteligencie je nutnou podmienkou pochopenie rozdielov medzi slabou a silnou, resp. medzi úzko špecializovanou a všeobecnou umelou inteligenciou v ich technologickej podstate. Rozsah diskutovaných tém môže byť skutočne rozsiahly – napr. sociálna spravodlivosť, ľudské práva, spravodlivá vojna, bioetické otázky, transhumanizmus, podstata ľudskej bytosti,...

Identifikácia oblastí, v ktorých sa seriózne nasadenie systémov umelej inteligencie nezaobíde bez uspokojivého návrhu riešenia etických problémov, zasa vyžaduje dôkladnú analýzu psychologických, sociologických i právnych aspektov ich nasadenia.

V horizonte vývoja všeobecnej umelej inteligencie môžu mať nezastupiteľnú úlohu viaceré

netechnické vedecké disciplíny, schopné prispieť k riešeniu modelov správania, teórie mysle, simulácie emócií i vedomia a pod.

K dôležitým rozmerom interdisciplinárnej komunikácie patrí aj spolupráca s inštitúciami na medzinárodnej i štátnej úrovni, bez ktorej nie je možné vytvárať reálne etické a právne regulačné rámce, a v neposlednom rade kooperácia s privátnym sektorom, ktorého vývojové kapacity a finančné i ľudské zdroje sú motorom vývoja a nasadenia systémov AI takmer do všetkých oblastí reálneho života.

4.3. Všeobecné návrhy

4.3.1. Umelá inteligencia zameraná na dobro človeka

Základným princípom pre akýkoľvek systém umelej inteligencie je zameranie na dobro človeka, teda známy a všeobecne prijímaný princíp **human-centered and beneficial artificial intelligence**. Navrhujeme však a zdôrazňujeme, že by princíp umelej inteligencie zameranej na človeka mal:

- **byť chápaný v duchu kresťanskej antropológie.**
- **zahŕňať každú ľudskú bytosť a nikoho nediskriminovať.**
- **mať na zreteli dobro ľudstva a spoločnosti, chrániac pri tom a rešpektujúc dobro každej ľudskej bytosti.**
- **sa vyznačovať starostlivosťou o náš „spoločný a zdieľaný domov“, teda o celý stvorený svet.**

Chápeme, že human-centered prístup zahŕňa aj mnohé technologické, praktické a právne náležitosti vývoja a nasadenia systémov umelej inteligencie (mnohé z nich sme už spomínali), no bez vyššie uvedeného návrhu aplikácia tohto princípu môže podliehať relativizmu a erózii hodnôt, či postupnému vyprázdneniu jeho podstatných aspektov.

Uchopenie princípu zamerania umelej inteligencie na človeka v duchu kresťanskej antropológie znamená prijať pohľad na človeka v kontexte Božieho zjavenia. V tejto perspektíve vidíme človeka, ktorý je stvorený na Boží obraz a pozvaný ku spáse. Vnímame realitu ľudského bytia, ktoré je zranené hriechom a v dimenziách pozemského putovania hľadá cestu k svojmu bytostnému naplneniu v Bohu. V perspektíve vedeckého bádania si uvedomujeme nielen všetko to, čo človeka na tejto ceste ovplyvňuje, ale i to, čo

sa stáva kontextom a prostriedkom tejto cesty napĺňanej v zmluvnom dialógu s Bohom a ostatnými ľuďmi. Prijímame radostnú zvesť o vykúpení a nachádzame vzor opravdivého človeka a skutočného naplnenia nášho človečenstva v Ježišovi Kristovi. Učíme sa nazerať na blížnych optikou evanjeliovej lásky a dobra s osobitným akcentom na službu a ochranu slabých a trpiacich. Povolaní do spoločenstva Najsvätejšej Trojice máme vytvárať ľudské spoločenstvo a budovať civilizáciu lásky.

Rámec kresťanskej antropológie a z neho prameniaci diapazón kresťanského humanizmu nám poskytuje aj dôležité morálne vyhranenie sa voči pokušeniu umelo vylepšovaného človeka prostredníctvom umelej inteligencie, či už na spôsob de Chardinovho vývoja od hominizácie cez divinizáciu až po bod Omega, alebo transhumanistického vylepšovania človeka, odovzdania primátu budúcej superinteligencii a pokriveného vnímania eschatologickej roviny viery v kontexte transfigurizmu, t.j. náboženského transhumanizmu.[493]

4.3.2. Dôveryhodná umelá inteligencia

Zameranie umelej inteligencie na človeka je základným predpokladom pre akýkoľvek systém umelej inteligencie, s ktorým môžeme bez zbytočných obáv interagovať a primerane mu dôverovať.

Principiálny postoj orientácie na človeka sa tak stáva ekvivalentným problematike dôveryhodnosti umelej inteligencie, pričom je treba stanoviť podmienky, bez splnenia ktorých by nasadenie systémov AI do reálneho sveta, v ktorom interagujú s človekom a vplývajú na spoločnosť, nemalo byť umožnené.

Na základe našich záverov z kapitol 2.8. a 3.1. i podľa Etického usmernenia pre dôveryhodnú umelú inteligenciu Skupiny expertov na umelú inteligenciu pri EÚ, ktoré bolo uvedené v kapitole 3.3., formulujeme **základné požiadavky na dôveryhodné systémy AI**. Tieto systémy musia byť:

- **legálne** (lawful) – vyhovovať požadovaným normám, zákonom i reguláciám

493 „Ide o celý komplex výziev a ponúkaných transhumanistických odpovedí, ktorých dôsledkom však je materialistický a technicistický redukcionizmus, ktorý redukuje všetko (pravdu, dobro, človeka a jeho mravné konanie) iba na technickú uskutočniteľnosť.“ S technokratickou redukciou svedomia tak smeruje k popieraniu ľudskej dôstojnosti v jej vnútornej integrite a k demontáži transcendentna, resp. redukcii ľudského bytia iba na technologickú úroveň.
VIVODA, *Transhumanizmus a Katolícka Cirkev*, s. 121-128.

a spĺňať všetky platné zákony a predpisy.

- **etické** (ethical) – rešpektovať etické zásady a hodnoty.
- **robustné** (robust) – dosahovať potrebné štandardy bezpečnosti a robustnosti[494] nielen z technologického hľadiska, ale zohľadňovať aj sociálne prostredie a dopady na spoločnosť.

Uvedomujeme si, že za každou z uvedených požiadaviek sa nachádza celý komplex problémov, procesov a dynamiky implementácie parciálnych riešení, o ktorých nemôžeme tvrdiť, že sú dokonalé a definitívne. Sú však cestou k cieľu, ktorým je dôveryhodná umelá inteligencia zameraná na človeka.

U každej z uvedených oblastí by sme mali definovať hranicu, od ktorej môžeme technológie umelej inteligencie považovať za dôveryhodné.

V oblasti noriem, zákonov a regulácií môžeme súhlasiť s aktuálnym obsahom pripravovanej regulácie Európskej únie v oblasti umelej inteligencie, Aktom o umelej inteligencii a považovať ho za primeranú legislatívnu hranicu pre súčasné dôveryhodné nasadenie systémov AI.

I napriek tomu, že Akt o umelej inteligencii v súčasnosti v rámci orgánov EÚ prechádza legislatívnym konaním, ktoré môže priniesť určité zmeny, dúfame, že dôležité regulačné princípy zostanú zachované, resp. vhodne doplnené o riešenie tých parciálnych problémov, ktoré by celoplošné nasadenie nariadenia diskvalifikovali. Len tak môže byť do platnosti uvedený dostatočne silný právny rámec dôveryhodnej umelej inteligencie, ktorý podľa vzoru GDPR ovplyvní veľkú časť sveta.

Minimálne etické princípy, zásady a hodnoty, ktoré by mali dôveryhodné technológie umelej inteligencie rešpektovať, uvádzame v nasledujúcej kapitole 4.3.3.

Ak odmyslíme všetky spoločenské problémy, ideologické a hodnotové rozdiely i legislatívne zápasy podstupované na poli umelej inteligencie, najproblematickejšou oblasťou pre dosiahnutie dôveryhodných systémov AI sa môže javiť **robustnosť, ktorú môžeme zjednodušene vyjadriť ako schopnosť bezpečne a spoľahlivo pracovať, resp. fungovať za akýchkoľvek podmienok**.

Dokonale robustný systém neexistuje a v kontexte súčasných moderných technológií si ani nevieme predstaviť jeho realizáciu v blízkej budúcnosti. Avšak na prevádzke jadrových

494 Ide o podchytenie a riešenie širokého rámca problémov, ktoré sme rozoberali v 2. kapitole.

zariadení, letoch do vesmíru, fungovaní viacerých oblastí kritickej infraštruktúry štátu, zabezpečení životných funkcií ľudí odkázaných na prístroje, atď. vidíme, že **vieme realizovať robustné systémy s primeranou mierou rizika**.

Minimálnu hranicu pre podmienku robustnosti dôveryhodného systému by sme mohli stanoviť takto:

- robustné systémy musia spĺňať vysoké technologické štandardy a normy, ktoré sú pre jednotlivé oblasti nasadenia záväzné.
- robustnosť je chápaná ako neustále prebiehajúci proces, v rámci ktorého sa schopnosti bezpečne a spoľahlivo pracovať vylepšujú, pričom sa existujúce chyby priebežne odhaľujú a opravujú.
- vývoj a prevádzka robustných systémov prebieha v rámci noriem riadenia kvality, manažmentu rizík a nahlasovania, serióznej analýzy a riešenia incidentov.

Treba zdôrazniť, že obsahom uvedených troch bodov je celý súbor opatrení a nariadení, pričom väčšina z nich je už implementovaná v európskom Akte o umelej inteligencii v časti tzv. významných povinností poskytovateľov, ktoré sme uviedli v kapitole 3.3. Ide o riadenie rizík, technickú dokumentáciu, interné záznamy (logovanie), technologickú transparentnosť, možnosť ľudského dohľadu a kvalifikovaného zásahu, prevádzkovú presnosť, spoľahlivosť a kybernetickú bezpečnosť, systém riadenia kvality, viaceré aspekty certifikácie (posúdenie zhody, registrácia, post-marketing monitoring) a oznamovanie incidentov.

Nielen podľa pripravovaného nariadenia EÚ, ale z podstaty veci sa k nim radí aj tzv. správa dát, požadujúca, aby testovacie a tréningové datasety boli „vysokej kvality“, aby tak systém AI, ktorý ich využíva, nebol diskriminačný a nevytváral nepredvídané, resp. nesprávne výsledky. Správa dát je však viacrozmerný problém, ktorý zasahuje do oblasti legálnosti, etiky i robustnosti dôveryhodných systémov AI.

Bez zodpovedného stanovenia a splnenia minimálnych požiadaviek v oblasti legálnosti, etiky a robustnosti systémov umelej inteligencie nie je možné hovoriť o dôveryhodnej umelej inteligencii.

4.3.3. Etické požiadavky na dôveryhodné systémy umelej inteligencie

V texte sme doteraz akcentovali viacero etických výziev, mnohokrát deklarovali etické zásady v rámci diskusie konkrétnej problematiky umelej inteligencie a sumarizovali aktuálne dianie v tejto oblasti. Na tomto pomerne širokom základe by sme chceli postaviť – podľa nás **zásadné – etické požiadavky na dôveryhodné systémy umelej inteligencie zameranej na človeka**:

- **pri vývoji, výrobe, nasadení, poskytovaní a používaní systémov umelej inteligencie musí byť zaručená ochrana slobody, dôstojnosti a bezpečia každej ľudskej osoby i celej spoločnosti.**
- **technológie umelej inteligencie musia byť plne pod ľudskou kontrolou a ovládateľné človekom.**
- **algoritmy i výsledky činnosti systémov AI musia byť človekom pochopiteľné a revidovateľné.**
- **akékoľvek nasadenie technológií AI musí byť prospešné pre človeka a spoločnosť.**
- **systémy umelej inteligencie nesmú byť nástrojom digitálneho rozdelenia.**
- **technológie umelej inteligencie nesmú škodiť nášmu spoločnému domu a mali by prispievať k spoločenskému a environmentálnemu blahobytu.**

Uznávame, že bez solídnej orientácie v problematike uvedené zásady nedokážeme dobre uchopiť a správne aplikovať, no považujeme ich za univerzálnejšie a viac principiálne, než konkrétne zásady, ktoré sme v rámci viacerých etických iniciatív predstavili.

Ak porovnáme štyri rôzne prístupy – závery vatikánskej konferencie renAIssance 2020 (Rome Call for Ethics), etické usmernenia skupiny expertov EÚ na umelú inteligenciu, sumarizované etické odporúčania z akademického sveta a etický základ pre tvorbu systémov umelej inteligencie od najväčšej organizácie zastrešujúcej moderné technologické normy a štandardy, pozorujeme rozmanitosť uchopenia základných etických pravidiel pre technológie umelej inteligencie. I keď sa v mnohom tematicky prelínajú, badať určité rozdiely prameniace z filozofických, sociologických, technologických, hodnotových a možno i ideologických predpokladov, na ktorých stavajú.

renAIssance 2020[495]	Ethics guidelines for trustworthy AI[496]	The global landscape of AI ethics guidelines[497]	IEEE Global Initiative on Ethics of Intelligent Systems[498]
transparentnosť	transparentnosť	transparentnosť	transparentnosť
zodpovednosť	zodpovednosť	zodpovednosť	zodpovednosť
inklúzia	rozmanitosť (diversity), nediskriminácia a spravodlivosť	spravodlivosť a férovosť	ľudské práva
nestrannosť			kompetencie
spoľahlivosť	technická odolnosť a bezpečnosť	neškodnosť	účinnosť
bezpečnosť a súkromie	bezpečnosť, ochrana súkromia a správa údajov	neškodnosť, súkromie	dátová agenda, informovanosť o zneužití
	spoločenský a environmentálny blahobyt		blahobyt

Máme za to, že náš prístup je univerzálnejší a môže byť nielen základom pre ktorúkoľvek z uvedených zásad jednotlivých iniciatív, resp. inštitúcií, ale dokáže i nadčasovo uchopiť budúci vývoj spoločnosti v konfrontácii a koexistencii s neustále sa rozvíjajúcim a technologicky napredujúcim fenoménom umelej inteligencie.

Realizácia (vývoj, výroba, nasadenie, poskytovanie, využívanie,...) systémov umelej

495 *Rome Call for AI Ethics (document)*. [on-line]. [cit. 28. marca 2022]. Dostupné na internete: <https://www.romecall.org/wp-content/uploads/2022/03/RomeCall_Paper_web.pdf>

496 *Ethics guidelines for trustworthy AI*. [on-line]. [cit. 28. marca 2022]. Dostupné na internete: <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>

497 *The global landscape of AI ethics guidelines*. [on-line]. [cit. 28. marca 2022]. Dostupné na internete: <https://doi.org/10.1038/s42256-019-0088-2>

498 *IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems*. [on-line]. [cit. 31. marca 2022]. Dostupné na internete: <https://standards.ieee.org/industry-connections/ec/autonomous-systems/>

inteligencie musí spĺňať:

- principiálny základ dôveryhodného systému umelej inteligencie zameraného na človeka.
- aplikáciu základných etických zásad vo všetkých rozmeroch, ktoré daný systém AI môže obnášať.
- splnenie legislatívnych požiadaviek.
- robustnú realizáciu a využívanie.

Aplikácia etických zásad a legislatívnych požiadaviek v reálnom svete by mala byť v rovine konkrétnych a jasných odporúčaní. Určite by sme sa voči našim zásadám neprehrešili, ak by sme stavali na kombinácii etických odporúčaní vatikánskej konferencie renAIssance 2020[499] a obsahu európskeho nariadenia Akt o umelej inteligencii.[500]

4.3.4. Oblasti implementácie etických princípov

V prehľade etických výziev, prameniacich z limitov a rizík súčasných systémov umelej inteligencie, ktoré sme priebežne uvádzali v 2. kapitole a sumarizovali v kapitole 3.1., nachádzame **tri oblasti nutnej implementácie etických noriem, eticko-právnych regulácií a morálnych zásad**:

- etické normy, zákonné regulácie a morálne zásady **tvorcov** systémov AI.
- etické normy, zákonné regulácie a morálne zásady **poskytovateľov a používateľov** týchto systémov.
- implementované eticko-právne požiadavky a obmedzenia priamo **v systémoch AI**.

Etický diskurz 4. kapitoly sa stále zameriava na technológie slabej umelej inteligencie (ANI)[501], u ktorých z podstaty veci musíme rátať so zaangažovaním ľudského faktora

499 *Rome Call for AI Ethics (document)*. [on-line]. [cit. 28. marca 2022]. Dostupné na internete: <https://www.romecall.org/wp-content/uploads/2022/03/RomeCall_Paper_web.pdf>

500 *Nariadenie Európskeho parlamentu a Rady, ktorým sa stanovujú harmonizované pravidlá v oblasti umelej inteligencie (Akt o umelej inteligencii) a menia niektoré legislatívne akty únie*. [on-line]. [cit. 24. marca 2022]. Dostupné na internete: <https://eur-lex.europa.eu/legal-content/SK/TXT/?uri=CELEX:52021PC0206>

501 Stále sa nachádzame v oblasti úzko špecializovaných systémoch umelej inteligencie (narrow AI), ktoré

vo všetkých oblastiach tvorby, používania i realizácie prostriedkov umelej inteligencie. Preto opakujeme jeden z našich záverov z kapitoly 3.1.: **aplikovanie etických princípov a regulácií v ANI – akokoľvek náročným až neuskutočniteľným by sa to v praxi mohlo ukázať – má vďaka prítomnosti ľudského faktora jasné zásady, vyhranené oblasti a viac menej presne definované kritériá**.

Treba preto investovať nemalé úsilie do vzdelávania, osvety i prevencie a formovať morálne postoje vývojárov, poskytovateľov i používateľov týchto technológií.[502]

A ako sme v kapitole 3.1. poznamenali, ruka v ruke s formáciou a vzdelávaním odborníkov v oblasti umelej inteligencie by sa mala rozvíjať aj **edukácia a osveta spoločnosti, bez ktorej si ťažko predstaviť rast spoločenskej citlivosti a zodpovednosti v oblasti celoplošnej adaptácie a využívania systémov AI**.

Osobitnou oblasťou je **základný výskum umelej inteligencie**, ktorý už z princípu nemôže byť mnohými reguláciami viazaný. I v tejto oblasti by však mali existovať etické garancie, ktorých cieľom je pri akomkoľvek type výskumu rešpektovať princíp zamerania na človeka a s tým súvisiacu dôveryhodnosť AI. Teda **Human-centered AI a Trustworthy AI považujeme za nutné princípy akéhokoľvek výskumu, vývoja, realizácie alebo nasadenia systémov umelej inteligencie**.

V kapitole 3.3. sme na margo technologických požiadaviek európskeho Aktu o umelej inteligencii diskutovali praktickú a v niektorých scenároch principiálnu nemožnosť splnenia eticko-právnych podmienok v technologickej rovine v prípade neskoršieho doplnenia technických úprav alebo procesných postupov do existujúcich systémov AI.

Ak v prípade kybernetickej bezpečnosti (NIS) alebo ochrany osobných údajov (GDPR) patria princípy *security by design* a *privacy by design* k tzv. osvedčeným a odporúčaným postupom (best practices), **v prípade systémov AI je *ethics by design* a *regulation by design* nutnou podmienkou ich vývoja, nasadenia a prevádzky**.[503] Etické zásady

sú optimalizované na zvládnutie konkrétnej úlohy, resp. množiny úloh. Ide súčasne o systémy slabej umelej inteligencie (weak AI), ktoré vykazujú inteligentné správanie na základe modelov a aplikovaných metód i tréningových dát. Sú to teda systémy zamerané na riešenie konkrétnych úloh a sú závislé na ľudskom vstupe a konfigurácii.

502 Napr. v predchádzajúcej kapitole spomínaná iniciatíva *IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems* zahŕňa i osobitné zameranie na etiku vývoja a vývojárov systémov AI.

503 Regulation by design v sebe obnáša aj security by design a privacy by design. Spolu s ethics by design tak tvoria celkový rámec trustworthy by design.

a právne normy tak musia byť integrálnou súčasťou prostriedkov umelej inteligencie už od ich technologického návrhu – neskorším doplnením sa v plnej miere a požadovanom rozsahu vo väčšine systémov AI nebudú dať spoľahlivo aplikovať.

Vzhľadom na autonómny a adaptívny rozmer technológií umelej inteligencie sa splnením princípov *ethics by design* a *regulation by design* kladie základ pre tzv. odolnosť týchto systémov voči budúcnosti (future-proof systems), t.j. schopnosť v budúcnosti spoľahlivo a bezpečne pracovať aj napriek možným nepredvídateľným okolnostiam.

4.4. Špecifické odporúčania pre algokraciu a armádne využitie

Popisujúc súčasné legislatívne aktivity v kapitole 3.3. sme s uspokojením konštatovali, že európsky Akt o umelej inteligencii má potenciál pokryť využívanie technológií umelej inteligencie prakticky vo všetkých oblastiach reálneho života spoločnosti. Predsa sme však spomenuli oblasti, v ktorých využitie algoritmov AI vybočuje z bežného rámca nasadenia. Ide o nasadenie systémov AI v oblasti pokročilého riadenia štátu, spravodajstva a plošného dohľadu a nasadenie vo vojenskej oblasti.[504]

Keďže ide o oblasti s veľkými právomocami, dosahom i rizikom zneužitia pre dobro človeka a celej spoločnosti, radi by sme upresnili, že do ich osobitnej regulácie by mala byť zahrnutá aj oblasť etických princípov a morálnych aspektov využívania.[505]

Odporúčania, ktoré uvádzame v tejto kapitole, by mali byť doplnkom k všeobecným návrhom a základným princípom, ktoré sme uviedli v kapitole predchádzajúcej.

4.4.1. Oblasť plošného dohľadu, spravodajstva a pokročilého riadenia štátu

Je pravdou, že niektoré z technológií algoritmického riadenia a plošného dohľadu pokrýva pripravované európske nariadenie Akt o umelej inteligencii v časti Zakázané systémy (čl.

504 Uvedomujeme si, že vzhľadom na osobitné riziká a výzvy existuje viacero oblastí so špecifickými etickými odporúčaniami a právnymi normami, napr. oblasť autonómnych vozidiel s ochranou ľudských životov, zabráneniu škodám a prisúdeniu zodpovednosti, bankový sektor s otázkou sociálnej spravodlivosti a transparentnosti, spoločnosť a mediálny svet s rizikami kapitalizmu dohľadu a manipulovania s ľuďmi, samostatné časti algoritmického riadenia so spravodlivosťou, transparentnosťou a demokratickými princípmi, prípadne rôznorodé kombinácie uvedených špecifík, napr. v zdravotníctve a pod.
Všetky tieto oblasti však dostatočne pokrýva navrhnutý Akt o umelej inteligencii v kombinácii s prípadnými špeciálnymi sektorovými reguláciami.

505 Etické princípy a morálne zásady v tejto oblasti ťažko odporúčať, treba ich vyžadovať.

5), no ide len o parciálne riešenie, ktoré je naviac sprevádzané možnými tzv. oprávnenými výnimkami. Ich legislatívne usmernenie musí byť preto pokryté inými zákonnými normami.

Opakujeme preto požiadavku, ktorú sme uviedli v kapitole 3.3.: **vzhľadom na špecifikum a dosah nasadenia systémov AI v oblasti pokročilého riadenia štátu, spravodajských služieb a plošného dohľadu[506] pre ľudské práva, ochranu demokracie a slobôd si myslíme, že táto oblasť by mala byť pokrytá už základnými legislatívnymi mechanizmami a verejným dohľadom demokratickej spoločnosti, ktoré sa týkajú riadenia spoločnosti, ľudských práv a pôsobenia spravodajských služieb vo všeobecnosti**.

Nie je jednoduché realizovať a vyžadovať striktnú zákonnosť i dôsledný dohľad demokraticky zvolených zástupcov a spoločnosti pri nasadení systémov AI v rámci spravodajských služieb, dohľadových systémov a všetkých foriem i stupňov algoritmického riadenia spoločnosti. Treba preto v celom spektre nasadenia od spravodajských služieb až po algoritmické riadenie akcentovať veľkú rozvážnosť a vyžadovať striktnú legálnosť i dôsledný dohľad demokraticky zvolených zástupcov, obmedziť dopady na sociálnu spravodlivosť a zabezpečiť dodržiavanie ľudských práv a hodnôt. Ide nielen o prijaté zákony, ale i nastavené procesy kontroly a mechanizmy zásahov v prípade podozrení na zlyhanie, či zneužitie činnosti technológií umelej inteligencie.

Osobitnou kapitolou je distribúcia pokročilých systémov umelej inteligencie do rizikových krajín. Podľa štúdie *The Global Expansion of AI Surveillance*, ktorú sme rozoberali v kapitole 2.6., je jednou zo znepokojujúcich skutočností fakt, že demokratické štáty, z ktorých väčšina sofistikovaných technológií algoritmického riadenia, spravodajstva a dohľadu pochádza, neprijímajú primerané opatrenia na monitorovanie a kontrolu šírenia týchto technológií, majúcich potenciál sa podieľať na celom rade možných porušení ľudských práv a zneužití autokratickými režimami a diktatúrami. Naviac je za ostatné desaťročie známych a zdokumentovaných viacero prípadov, keď sa technologické spoločnosti z demokratických krajín priamo podieľali na nasadení svojich technológií v scenároch obmedzujúcich ľudské práva, slobodu, súkromie a demokraciu.[507] Len pomaly rastie celospoločenské povedomie a angažovanie pri obmedzovaní exportu do rizikových

506 Viaceré aspekty nasadenia systémov AI v oblasti spravodajstva a algokracie sme diskutovali v kapitole 2.6.

507 Napr. Microsoft, Cisco a ďalší pri sledovacích systémoch v Číne....

krajín.

Navrhujeme, aby oblasť exportu produktov a technológií umelej inteligencie, ktoré môžu byť zneužité v oblasti pokročilého riadenia štátu, spravodajstva a plošného dohľadu, bola predmetom medzinárodnej regulácie s cieľom zamedziť ich vývoz do rizikových krajín. Sankcie by však nemali byť nástrojom (geo) politických zápasov, ale skutočného nasadenia na poli etického využívania systémov AI.

4.4.2. Systémy umelej inteligencie vo vojenskej oblasti

V kapitole 3.1. sme ako jeden zo záverov rizík nasadenia technológií umelej inteligencie vo vojenskej oblasti uvádzali: „Schopnosti umelou inteligenciou poháňaných autonómnych zbraňových systémov a kybernetických zbraní i napriek všetkým rizikám vedú k zvyšujúcim sa tlakom na financovanie a zavádzanie útočných kybernetických zbraní. Jednotlivé krajiny sa nedokážu vzdať tak lákavej technologickej výhody a sú pevne rozhodnuté technológie umelej inteligencie implementovať v celej šírke možného zmysluplného využitia. Problematika obmedzenia týchto útočných systémov je skomplikovaná aj reálnym stieraním hraníc medzi obranným a útočným nasadením takmer vo všetkých oblastiach vojenského využitia technológií umelej inteligencie.“

Túto skutočnosť sme následne v kontexte prijímaných regulácií v kapitole 3.3. komentovali slovami: „Keďže akékoľvek obmedzovanie technológií umelej inteligencie vo vojenskej oblasti môže byť chápané ako bezpečnostné riziko a zníženie bojaschopnosti modernej armády, jednostranne prijaté regulácie nemusia byť účinné – nielen pre to, že sa jednou stranou ťažko prijímajú (aj keď pre hodnotovo orientovanú spoločnosť by to malo byť povinnosťou), ale i pre malú šancu na ich extra teritoriálne rozšírenie a akceptovanie.“[508]

Mnohí štátni aktéri by chceli aplikovať latinské *si vis pacem, para bellum* i na oblasť armádneho nasadenia technológií AI v zbraňových systémoch a ich potenciálom upevňovať svoje postavenie. Pre ďalších – pozerajúc k horizontu možností autonómnych zbraňových systémov – by sa ich zavádzanie do výzbroje mohlo podobať odstrašujúcemu potenciálu jadrových zbraní. Ak však pozeráme za horizont súčasných možností vojenských systémov AI, vidíme technológie, ktorých rizikový potenciál prekračuje nebezpečenstvo prameniace z terajších arzenálov jadrových zbraní.

Rizikám a limitom systémov umelej inteligencie vo vojenskom nasadení a v zbraňových

508 Kapitola 3.3.

systémoch sme sa obšírne venovali v kapitole 2.7., pričom ich etické dôsledky sme diskutovali v kapitolách 3.1. a 3.2.

V otvorenom liste *Autonomous Weapons: An Open Letter from AI & Robotics Researchers*, ktorý sme spomínali v kapitolách 2.7.8. a 3.2. a ktorý vyzýval na zákaz autonómnych zbraňových systémov, sú tieto systémy označované ako tretia revolúcia v zbraňových systémoch, po strelnom prachu a jadrových zbraniach. Ide o posun nielen v technológii, ale – a to je pre armádne nasadenie dôležité – predovšetkým o posun v schopnostiach a účinnosti týchto zbraní, ktorých plné nasadenie by mohlo viesť až k nukleárnej, či jej podobnej katastrofe.

Nadväzujúc tak na Russell-Einsteinov manifest, varujúci pred reálne možnými rizikami jadrového zbrojenia, pripájame sa k všetkým výzvam, ktoré volajú po zákaze smrtiacich autonómnych zbraňových systémov. Či už išlo o *Scientists' Call to Ban Autonomous Lethal Robots* z roku 2013, spomínaný *Autonomous Weapons: An Open Letter from AI & Robotics Researchers* z roku 2015, v roku 2017 vydaný *An Open Letter to the United Nations: Convention on Certain Conventional Weapons*, alebo európske *Resolution on autonomous weapon systems* z roku 2018.[509]

Principiálny postoj v oblasti smrtiacich autonómnych zbraňových systémov (LAWs) je jasný: **technológiami umelej inteligencie poháňané automatické smrtiace zbraňové systémy, systémy automatického zameriavania a vyberania cieľov, automatické systémy schopné bez zásahu človeka rozhodnúť o smrtiacej reakcii akéhokoľvek druhu (od útoku dronu až po rozpútanie jadrového konfliktu) musia byť zakázané**.

Rozoberajúc limity a riziká súčasných technológií umelej inteligencie,[510] ani pri najlepšej vôli nie sme schopní vytvoriť také systémy ANI, ktorým by sme mohli zveriť samostatné rozhodovanie v tak dôležitej oblasti. A pokiaľ ide o AGI – znovu sa opakujúc – ak nemáme vyriešené otázky ako napr. funkčný model správania, implementáciu komparatívnych hodnotových rámcov a pod., nie sme schopní o tejto možnosti ani len špekulatívne uvažovať.

Vývoj a nasadenie viacerých extrémne nebezpečných vojenských technológií je v súčasnosti na základe vzájomného konsenzu a právnych záväzkov zakázané. Je preto žiadúce, aby sa tak stalo aj v prípade technológiami umelej inteligencie poháňaných plne

509 Všetky uvedené iniciatívy sme diskutovali v kapitole 3.2.

510 Znovu sa odvolávame na kapitoly 2.1. až 2.4.

automatizovaných smrtiacich zbraňových systémov, systémov automatického zameriavania a vyberania cieľov i automatických systémov schopných bez zásahu človeka rozhodnúť o smrtiacej reakcii akéhokoľvek druhu.

Od plne autonómnych zbraňových systémov bez kontroly človeka sa však vráťme k tomu, čím sme túto kapitolu začali – strategickému i praktickému zavádzaniu takých technológií AI do armádnych systémov, ktoré by mali byť pod kontrolou človeka.

V kapitolách 2.7.8. a 3.2. spomenuté eticko-právne procesy, ktoré prebiehajú napríklad v armáde USA, môžu byť základom pre celosvetovú diskusiu mocností a na základe tlaku verejnosti, angažovanosti jednotlivých častí spoločnosti v rôznych regiónoch sveta i úsilia zodpovedných strán môžu viesť k prijatiu celosvetových pravidiel i záväzkov pre oblasť vývoja a nasadenia rizikových vojenských systémov vybavených technológiami umelej inteligencie.

Ako základ môže poslúžiť v kapitole 2.7.8. diskutovaný dokument *AI Principles: Recommendations on the Ethical Use of Artificial Intelligence* by the Department of Defense z Defense Innovation Board USA s podstatnými závermi, ku ktorým patrí zodpovednosť (rozhodovacia právomoc), opatrnosť pri príprave testovacích dát a návrhu systému, dosledovateľnosť, spoľahlivosť a ovládateľnosť.[511]

Okrem väčšiny všeobecných požiadaviek, ktoré sme v kapitole 4.3. uvádzali a ktoré je možné vo vojenskom využití aplikovať, mali by v prípade autonómnych zbraňových systémov kontrolovateľných človekom platiť nasledovné zásady:

- **nutnou podmienkou prevádzky ľubovoľného systému AI, ktorý môže predstavovať riziko pre akúkoľvek ľudskú osobu, je schopnosť a možnosť človeka prebrať kedykoľvek kontrolu nad týmto systémom, resp. právo a možnosť verifikovať a prehodnotiť výsledky jeho činnosti**.
- **limity, regulácia a obmedzenia LAWs by mali predstavovať etický rámec stanovený na základe morálnych hodnôt ľudskej spoločnosti, nie na základe relativistickej tzv. „následnej regulácie“.[512]**

Vráťme sa k rôznorodosti štátnych aktérov a armádneho zápolenia, ktoré sme v úvode

511 Vidíme, že ide o dosť okliešťený regulačný rámec oproti iniciatívam, ktoré sme rozoberali v kapitole 4.3.

512 Fenomén „následnej regulácie“, ktorá smeruje k hodnotovému a morálnemu relativizmu a snaží sa prehodnotiť argumenty o nenahraditeľnosti ľudského svedomia a morálneho úsudku, sme diskutovali v kapitole 2.7.8.

tejto kapitoly spomínali a ktoré pomerne efektívne bráni v možnom jednostrannom, resp. dobrovoľnom obmedzovaní zavádzania rizikových zbraňových systémov AI.

Uviedli sme, že napriek zníženiu konkurenčnej schopnosti a malej šanci na akceptovanie inými štátmi, by jednostranne prijaté (vyššie uvedené) eticko-právne regulácie mali byť pre hodnotovo orientovanú spoločnosť povinnosťou. Pozývame k tomuto kroku z viacerých dôvodov:

- žiadny štát, pokiaľ sa zriekne etických princípov a morálnych zásad, nemá právo obhájiť svoju účasť na vojnovom konflikte a zvíťaziť.
- **pri nasadení moderných zbraňových systémov s celoplošnými účinkami a technológií, zasahujúcich v hybridných vojnách a vojnách 4. generácie prakticky celé populácie štátov, sa koncept spravodlivej vojny stáva neprijateľný**.
- len na základe konkrétne prijatých záväzkov sa môže celosvetová diskusia mocností, tlak verejnosti, angažovanosť jednotlivých častí spoločnosti v rôznych regiónoch sveta a úsilie zodpovedných strán pretaviť v postupné prijatie celosvetových pravidiel i záväzkov pre oblasť vývoja a nasadenia rizikových vojenských systémov vybavených technológiami umelej inteligencie.

V čase písania tejto kapitoly nás používanie viacerých sofistikovaných útočných bojových systémov v prebiehajúcom vojnovom konflikte na Ukrajine stavia pred dilemu, či budeme technológie umelej inteligencie využívať v duchu hesla *si vis pacem, para* bellum vo falošnej predstave, že vďaka nim sa vyhneme vojnovým konfliktom, pričom však budeme mať tendenciu ich v potencionálnych konfliktoch aj použiť.

Alebo v intenciách *si vis pacem para pacem* – vyvarujúc sa pacifistickému nevyužitiu ich potenciálu – dokážeme byť dostatočne zrelí na ich obranné využitie a nasadenie pre ochranu života a ľudskej dôstojnosti i zabezpečenie skutočných hodnôt spoločnosti.

4.5. Priestor pre Cirkev – angažmán, ktoré treba prijať

4.5.1. Morálno-etický diskurz a misia zjednocovať, usmerňovať i propagovať etické aktivity

V kapitole 3.4. sme sa v kontexte štúdie *The global landscape of AI ethics guidelines* pozastavovali nad celkovou hodnotovou a etickou diskrepanciou vo svete: v zásade

panuje celospoločenský konsenzus ohľadom potreby etických usmernení v oblasti umelej inteligencie, tento je však sprevádzaný podstatnými rozdielmi v chápaní etiky a hodnôt.[513]

Pripomeňme si preto slová Mons. Vincenza Pagliu na margo rímskej konferencie renAIssance 2020: „Zámerom výzvy je vytvoriť hnutie, ktoré sa rozšíri a zapojí ďalších aktérov: verejné inštitúcie, mimovládne organizácie, priemyselné odvetvia a skupiny, aby určili smer vývoja a používania technológií odvodených od umelej inteligencie. Z tohto hľadiska môžeme povedať, že prvý podpis tejto výzvy nie je vyvrcholením, ale východiskom pre záväzok, ktorý sa javí ako ešte naliehavejší a dôležitejší než kedykoľvek predtým. Pripojenie sa k tejto iniciatíve pre zástupcov priemyslu, ktorí ju podpíšu, znamená záväzok, ktorý má význam aj z hľadiska nákladov a technologického príspevku k vývoju a distribúcii ich výrobkov. Ak sa Akadémia cíti byť povolaná zintenzívniť svoje úsilie o uľahčenie získavania poznatkov a podpisov iných medzinárodných aktérov, táto výzva je len prvým krokom, po ktorom by mali nasledovať ďalšie. Text výzvy sa vyznačuje aj tým, že je prvým pokusom sformulovať súbor etických kritérií so spoločnými referenčnými bodmi a hodnotami, čím ponúka príspevok k rozvoju spoločného jazyka na interpretáciu toho, čo je človek“.[514]

Katolícka cirkev má v rámci svojho pôsobenia a univerzálneho spoločenstva výnimočnú možnosť **iniciovať a rozvíjať celosvetové širokospektrálne hnutie s cieľom etického vývoja a používania technológií umelej inteligencie**.

Na základe Zjavenia, solídneho teologického aparátu a interdisciplinárnych skúseností má potenciál **identifikovať, uchopiť, rozpracovať a formulovať základný univerzálny súbor etických kritérií a hodnôt** ako odpoveď na etický a hodnotový relativizmus.[515]

---

513 Por. JOBIN, IENCA, VAYENA, *The global landscape of AI ethics guidelines*. [on-line]. [cit. 28. marca 2022].
Dostupné na internete: <https://doi.org/10.1038/s42256-019-0088-2>

514 *Rome Call for AI Ethics*. [on-line]. [cit. 19. augusta 2020].
Dostupné na internete: <http://www.academyforlife.va/content/pav/en/events/intelligenza-artificiale.html>

515 Viď napr. „Etické rozhodovanie vo všeobecnosti neprináša absolútne správne alebo absolútne nesprávne riešenia daných situácii. Inými slovami, rozhodnutie, ktoré istá skupina ľudí považuje za správne, resp. v ich ponímaní etické, nemusí automaticky predstavovať správne (etické) rozhodnutie v očiach iných ľudí.“
ŠTARHA, GAŠPAROVIČ, *AI z pohľadu práva*. [on-line]. [cit. 29. marca 2022].
Dostupné na internete: <https://www.epravo.sk/top/clanky/ai-z-pohladu-prava-4483.html>

Vďaka svojej angažovanosti od diplomatických misií až po periférie sveta má možnosť **ponúkať a inkulturovať spoločné referenčné body a hodnoty základných etických kritérií do rôznorodosti ľudského spoločenstva a jednotlivých oblastí nasadenia systémov AI v spoločnosti**.

Predstavitelia Cirkvi môžu tiež pôsobiť ako **neutrálni sprostredkovatelia v špecifických oblastiach**, napr. v rámci jednaní o obmedzení vojenského nasadenia technológií umelej inteligencie a pod.

Keďže vplyv a dôsledky využívania AI budú viac a viac zasahovať prakticky do všetkých oblastí života dnešného sveta, musí byť táto oblasť z pohľadu etického využívania predmetom neustálej osvety a angažovanosti zo strany Cirkvi – podobne, ako sa deje v prípade viacerých dôležitých aktivít na poli OSN, medzinárodnej úrovni i každodennej práce Cirkvi v dnešnej spoločnosti.

4.5.2. Akcent na univerzálne bratstvo a sociálne priateľstvo

V duchu encykliky pápeža Františka *Fratelli tutti – o bratstve a sociálnom priateľstve* je Cirkev pozvaná prispieť k využívaniu technológií umelej inteligencie pre dobro celého ľudského spoločenstva na zemi: „...moja kritika technokratickej paradigmy neznamená, že môžeme zostať v bezpečí, len ak sa budeme usilovať kontrolovať jej excesy. Najväčšie nebezpečenstvo nespočíva často vo veciach, v materiálnych skutočnostiach, organizáciách, ale v spôsobe, akým ich ľudia používajú.“[516]

Myšlienka budovať univerzálne bratstvo a sociálne priateľstvo, ktorá rezonovala naprieč tzv. oblasťami vplyvu definovanými v záveroch rímskej konferencie renAIssance 2020, je mnohorakým spôsobom uchopená a rozvinutá v encyklike Fratelli tutti.

V rámci týchto dokumentov ako príklad uvádzame niekoľko výziev, v ktorých – okrem všetkého na poli etiky doteraz uvedeného – má Cirkev potenciál predstavovať fenomén umelej inteligencie ako nástroj účasti na evanjeliovom budovaní univerzálneho bratstva a sociálneho priateľstva:

- **premostenie k marginalizovaným** (spomínaný *bridge the gap* a zmierňovanie *digital divide*).

516 PÁPEŽ FRANTIŠEK. *Fratelli tutti*. [on-line]. Čl. 166. [cit. 4. apríla 2022]. Dostupné na internete: <https://www.kbs.sk/obsah/sekcia/h/dokumenty-a-vyhlasenia/p/dokumenty-papezov/c/encyklika-fratelii-tutti>

- **nenechávať nikoho za sebou** (nemo resideo, no one left behind).
- vedomie, že **ak chceme budovať nejakú budúcnosť, musíme ju budovať všetci**, inak sa to nedá.

Sme pozvaní podstúpiť zápas o pravú tvár etiky umelej inteligencie – aby sa z nej nestal nástroj moderných ideológií, sociálnej nespravodlivosti, obmedzovania ľudských práv, digitálneho rozdelenia, erózie hodnôt a ohrozenia ľudského bytia i celého spoločenstva – ale aby bola dobrým sluhom, prostriedkom budovania univerzálneho bratstva a sociálneho priateľstva.[517]

517 Ide o jeden zo zámerov vyjadrených v úvode tohto diela – technológie umelej inteligencie v úlohe dobrého sluhu, nie zlého pána.

Obr. č. 12. Návrh riešenia etických požiadaviek na dôveryhodné systémy AI.

5. Vízia silnej a všeobecnej umelej inteligencie

Rozhodli se svěřit naše bezpečí něčemu, co není schopné věrnosti, morálky ani moudrosti.[518]

Akokoľvek je táto publikácia primárne zameraná na súčasné technológie umelej inteligencie, ktoré sa napriek mnohorakej mediálnej nadsádzke nedajú nazvať inak než len slabými systémami ANI, prakticky pri všetkých doteraz rozoberaných témach sme sa nevyhli aspoň krátkemu pohľadu za horizont – k systémom silnej a uvedomelej umelej inteligencie, ktorú síce mnohí s nádejou očakávajú, no iní sa jej obávajú. Pritom však všetci stoja pred dilemou, ako ju uchopiť...

Protagonisti Dartmouthského seminára, ktorý znamenal zrod fenoménu umelej inteligencie,[519] si kládli za cieľ „pokúsiť sa zistiť, ako prinútiť stroje používať jazyk, vytvárať abstrakcie a pojmy, riešiť druhy problémov, ktoré sú teraz vyhradené pre ľudí, a zlepšovať sa.“[520]

Účastníci seminára mali veľký entuziazmus a optimizmus ohľadom dosiahnutia pokročilej umelej inteligencie, čo John McCarthy vyjadril slovami: „myslíme si, že významný pokrok v riešení jedného alebo viacerých z týchto problémov sa dá dosiahnuť, ak starostlivo vybraná skupina vedcov bude na tom spoločne pracovať počas jedného leta“.[521]

Vôbec sa netajili zámerom svoje vízie a snahy zamerať na realizáciu uvedomelej umelej inteligencie.

Postupom času – napriek všetkému vynaloženému úsiliu – prichádza vytriezvenie. McCarthy na začiatku šesťdesiatych rokov zakladá Standfordský projekt umelej inteligencie „s cieľom vytvorenia plne inteligentného stroja v rámci dekády“.[522] Budúci

518 CAMPBELL, J. *Dvojník.* Fantom Print 2015, ISBN 978-80-7398-329-7, s. 713.

519 Dartmouthský seminár, ktorý sa uskutočnil v r. 1955, sme uvádzali v kapitole 1.1.

520 MITCHELL, M. *Conceptual Abstraction and Analogy in Artificial Intelligence*. In: *ALIFE 2020: The 2020 Conference on Artificial Life*. [on-line]. [cit. 5. apríla 2022].
Dostupné na internete: <https://doi.org/10.1162/isal_a_00354>

521 MCCARTHY et al., *Proposal for the Dartmouth Summer Research Project in Artificial Intelligence*. [on-line]. [cit. 3. februára 2021].
Dostupné na internete: <https://doi.org/10.1609/aimag.v27i4.1904>

522 MORAVEC, *Mind Children: The Future of Robot and Human Intelligence*, s. 20.

laureát Nobelovej ceny v tom istom čase predikuje: „Do dvadsať rokov budú stroje schopné vykonávať akúkoľvek prácu, ktorú môže vykonávať človek“.[523] A Minský obnovuje nádeje, hovoriac: „v rámci jednej generácie budú problémy spojené s vytvorením umelej inteligencie v podstate vyriešené“.[524]

Pomaly dobiehame siedme desaťročie od Dartmouthského seminára a napriek úžasnému pokroku – osobitne v posledných dekádach – žiadna z týchto predpovedí sa doteraz nenaplnila. Výskumné témy, predstavené na seminári, sú stále otvorené a aktívne skúmané v odbornej komunite po celom svete.

Hoci umelá inteligencia dosiahla za posledné desaťročie dramatický (dá sa povedať, že parciálne až exponenciálny) pokrok v oblastiach, ako je videnie, spracovanie prirodzeného jazyka a robotika, súčasným systémom AI stále takmer úplne chýba schopnosť vytvárať pojmy a abstrakcie podobné ľudským, schopnosť porozumieť, chápať súvislosti a myslieť, realizovať niečo na spôsob zdravého rozumu ako predpokladu pre všeobecnú inteligenciu.[525]

5.1. „Trhliny v inteligencii“ pokročilých systémov AI

Méta uvedomelej umelej inteligencie by sa mala dosiahnuť prostredníctvom silnej (strong) a všeobecnej (general) AI. Všeobecnej, lebo dokáže zvládnuť akúkoľvek intelektuálnu úlohu a má schopnosť generalizovať, t.j. zovšeobecňovať a prenášať, či adaptovať naučené schopnosti na iné úlohy. Silnej, pretože aj skutočne rozumie tomu, čo rieši a vykonáva.

Podľa protagonistov uvedomelej AGI by na tomto základe mal byť možný rozvoj umelej

523 SIMON, H. A. *The Shape of Automation for Men and Management*. New York: Harper & Row, 1965, s. 90.

524 MINSKY, M. L. *Computation: Finite and Infinite Machines*. Upper Saddle River, N.J.: Prentice Hall, 1967, s. 2.

525 Gary Marcus, profesor psychológie a špecialista v oblasti umelej inteligencie uvádza: pri vývoji „silnej umelej inteligencie“ – teda všeobecnej umelej inteligencie na úrovni človeka – „nie je takmer žiadny pokrok“.
PRESS, G. *12 Observations About Artificial Intelligence From The O'Reilly AI Conference*. In: *Forbes.* [on-line]. 2016, 31. októbra. [cit. 7. augusta 2020].
Dostupné na internete: <https://www.forbes.com/sites/gilpress/2016/10/31/12-observations-about-artificial-intelligence-from-the-oreilly-ai-conference/>

inteligencie, ktorá by bola porovnateľná (ak nie lepšia) s mysľou človeka, smerujúc tak k vedomiu a sebauvedomeniu, až po analógiu ľudskej osoby.

Nech je však rozvoj súčasných systémov umelej inteligencie akokoľvek prevratný a dosiahnuté výsledky impozantné, stále narážame na podstatné problémy, bez ktorých sa o skutočnej AGI nedá uvažovať.

Predovšetkým ide o tzv. **bariéru chápania zmyslu** (barrier of meaning) medzi súčasnými systémami AI a inteligenciou na ľudskej úrovni.[526] Algoritmy ANI podstatným spôsobom nedokážu chápať zmysel, ani sa vyrovnať ľuďom vo vnímaní, reči a chápaní. Systémy ANI preto konajú chyby diametrálne odlišné od tých ľudských a majú problémy s abstrahovaním[527] a prenášaním skúseností, resp. toho, čo sa naučili. Technológiám ANI chýba chápanie zdravého rozumu, keďže **miesto myslenia realizujú jeho simuláciu**. Rozdiel medzi myslením a jeho napodobnením v pochopení zmyslu, súvislostí, analógií,[528] konceptov a mnohých ďalších súvislostí je tak veľký, že súčasné systémy ANI sú náchylné na jednoduché oklamanie na základe reálneho nechápania obsahu, súvislostí a zmyslu.[529]

V rámci systémov AI nie sme schopní vytvoriť niektoré základné prvky analogické ľudskému chápaniu, ktoré máme či už vrodené, alebo sa vyvíjajúce v rámci prvých rokov života. Viacerí psychológovia hovoria napríklad o tzv. **intuitívnej fyzike** – základnom, rýchlo osvojenom a zovšeobecnenom poznaní o objektoch hmotného sveta, **intuitívnej**

526 MITCHELL, *Artificial Intelligence*, s. 235.

527 Už v rámci propozícií Dartmouthského seminára sa „formovanie abstrakcií“ uvádza ako jedna z kľúčových schopností umelej inteligencie. Ako však uvádza prof. Melanie Mitchell, **schopnosť systémov AI formovať ľudskému mysleniu sa podobajúce konceptuálne abstrakcie zostáva až doteraz takmer vôbec nevyriešená**.

528 Niektorí kognitívni vedci navrhli, že tvorba analógií je ústredným mechanizmom pojmovej abstrakcie a porozumenia u ľudí. Douglas Hofstadter, ktorého sme spomínali v úvode tejto publikácie, nazval tvorbu analógií „jadrom poznávania“, pričom so spoluautorom Emmanuelom Sanderom poznamemáva: „Bez pojmov nemôže byť myslenie a bez analógií nemôžu byť pojmy.“ Schopnosti tvorby analógií sú kľúčové pri vývoji systémov umelej inteligencie s inteligenciou podobnou ľudskej.
MITCHELL, *Conceptual Abstraction and Analogy in Artificial Intelligence*. [on-line]. [cit. 5. apríla 2022]. Dostupné na internete: <https://doi.org/10.1162/isal_a_00354>

529 Práve **chyby nepodobné ľudským, ktorých sa systémy AI dopúšťajú a ktoré sme v tejto publikácii viackrát uvádzali (napr. zraniteľnosti voči tzv. adversarial examples), poukazujú na nepochopenie konceptov nachádzajúcich sa vo vstupných dátach**. Systémy AI sa pri vhodnej voľbe algoritmov, hyperparametrov a tréningových dát dokážu vytrénovať na úzko špecializované úlohy, no nedokážu chápať princípy, koncepty a zmysel toho, čo vykonávajú.

biológii – uvedomovaní si rozdielov medzi živými a neživými objektami, alebo o **intuitívnej psychológií** – sociálnej schopnosti vnímať pocity, postoje a zámery iných osôb. U človeka pritom ide len o fundamentálnu bázu poznatkov, na základe ktorých sa rozvíjajú kognitívne schopnosti, rôznorodé aspekty učenia sa a myslenia, ku ktorým patrí schopnosť spoznávať nové koncepty len z niekoľkých príkladov, schopnosť ich zovšeobecňovať a aplikovať na iné situácie, rýchlo chápať zmysel reálnych situácií a na ich základe sa prakticky okamžite, adekvátne a správne rozhodovať.[530]

To, čo je v rámci intuitívnej fyziky, biológie a psychológie bežnou súčasťou prvých rokov rozvoja ľudskej osobnosti, dokážu v súčasnosti zvládnuť najlepšie systémy AI len vo veľmi obmedzenom a neúplnom podaní. Naviac k tomu potrebujú extrémne množstvo vstupných dát, trénovacích cyklov a vysoký výpočtový výkon. A akúkoľvek pokročilú oblasť, v ktorej sú technológie AI dokonca lepšie než ľudia, vieme dosiahnuť len za cenu extrémnej špecializácie a úzkeho zamerania systémov AI.[531]

I to odráža **súčasnú neexistenciu modelu mysle a modelovania správania systémov AI**, keďže algoritmy umelej inteligencie nie sú schopné na základe skutočného chápania vytvárať tzv. mentálne modely dôležitých aspektov reálneho života a sveta. Ide v zásade o mentálne simulácie fungovania sveta, ktoré sa nám premietajú v mysli a na základe ktorých vieme nielen očakávať, čo sa stane, ale sa aj zmysluplne rozhodovať.[532]

Algoritmy AI, i napriek veľkým pokrokom v rozpoznávaní reči a snahy strojovo pochopiť význam slovných vyjadrení, nedokážu uchopiť metafory, ktorými sa bežne vyjadrujeme, keďže – ako sme už viac krát uviedli – majú problém s abstrakciou a analógiou. Ľudská myseľ abstrahuje a vytvára analógie (vo väčšine prípadov podvedome) a tvorí tak základ pre formovanie konceptov. A koncepty sú základom chápania i celkového ľudského

---

530 Por. MITCHELL, *Artificial Intelligence*, s. 236-237.
Por. PEARL, J., MACKENZIE, D. *The Book of Why: THe New Science of Cause and Effect*. New York: Basic Books, 2018.

531 Napr. v súčasnosti asi najsofistikovanejšie systémy umelej inteligencie v oblasti hier Go a šach, ktorými sú viaceré verzie AlphaGo a AlphaGo Zero, sú absolútne nepoužiteľné napr. v autonómnych vozidlách, inteligentných zbraňových systémoch, onkologických alebo astrofyzikálnych systémoch a pod.
Inak povedané, špičkové úspechy konkrétnych algoritmov AI je v súčasnosti možné dosiahnuť za dosť obmedzených podmienok, tzv. typov charakteristík ovplyvňujúcich činnosť inteligentného agenta.
Por. RUSSELL, *Human Compatible*, s. 44, 56.

532 Por. MITCHELL, *Artificial Intelligence*, s. 240-242.

modelu myslenia a schopnosti adekvátne konať, resp. reagovať na situácie s ktorými sa človek v živote stretáva.[533]

Schopnosť abstrakcie, analógie, tvorby konceptov a využívania metafor patrí k nutným podmienkam pre vytvorenie modelu myslenia založeného na zdravom rozume. Pokiaľ systém umelej inteligencie nebude toho schopný, nikdy neprekročí hranicu medzi slabou a silnou umelou inteligenciou. Mnohí odborníci z oblasti umelej inteligencie sa zamýšľajú nad otázkou, či tieto tzv. „**trhliny v inteligencii**“ (cracks in intelligence) systémov AI môžu byť opravené vylepšovaním algoritmov a prísunom ďalších dát, alebo či niečo podstatné nám v týchto systémoch chýba.[534]

Už od čias publikovania článku Alana Turinga o univerzálnom počítacom stroji, ktorý mal byť schopný vyriešiť ľubovoľnú výpočtovú a logickú operáciu (Turingov stroj[535]) badať snahu vnímať schopnosť vykonávať logické operácie ako niečo samostatné, oddelené od subjektu, ktorým bola do tej doby výhradne ľudská bytosť. Túto ideu, siahajúcu až k Descartesovmu špekulatívnemu „naše telá a naše myšlienky sa skladajú z rôznych látok a podliehajú rôznym fyzikálnym zákonom“[536], si naplno osvojili aj účastníci Dartmouthského seminára. Dlhodobo (a dodnes) dominantným prístupom sa tak stalo presvedčenie, že všeobecná a silná umelá inteligencia môže byť dosiahnutá od ľudského tela oslobodenými počítačmi.

V komunite venujúcej sa umelej inteligencii však stále pretrvávala aspoň malá skupina vedcov, ktorá zastávala tzv. **hypotézu stelesnenia** (embodiment hypothesis) založenú na premise, že stroj nemôže dosiahnuť inteligenciu ľudskej úrovne bez nejakého druhu „tela“, v ktorom interaguje s okolitým svetom, pretože mozog oddelený od tela nikdy

---

533 HOFSTADTER, D. R. *Analogy as the Core of Cognition*. [on-line]. Presidential Lecture, Stanford University, 2009. [cit. 7. apríla 2022].
Dostupné na internete: <https://www.youtube.com/watch?v=n8m7lFQ3njk>
HOFSTADTER, D. R, SANDER, E. *Surfaces and Essences*. New York: Basic Books, 2013, s. 3.

534 MARCUS, G. *Deep Learning: A Critical Appraisal*. [on-line]. [cit. 7. apríla 2022].
Dostupné na internete: <https://arxiv.org/abs/1801.00631>

535 *Turing machine*. [on-line]. [cit. 7. apríla 2022].
Dostupné na internete: <https://en.wikipedia.org/wiki/Turing_machine>

536 *Dualism*. [on-line]. [cit. 7. apríla 2022].
Dostupné na internete: <https://plato.stanford.edu/entries/dualism/>

nemôže nadobudnúť koncepty potrebné pre všeobecnú inteligenciu.[537]

V súčasnosti, vnímajúc vyššie opísané fundamentálne problémy a stagnáciu v smerovaní k silnej a všeobecnej umelej inteligencii, sa začína čím ďalej tým viac odborníkov prikláňať k hypotéze stelesnenia a veľa z nich pretavilo tento posun do svojej zaangažovanosti vo vývoji systémov, ktoré interagujú s reálnym svetom (napr. autonómne vozidlá alebo robotické systémy, ktorých plná autonómnosť a akcieschopnosť sa viaže na užívanie zdravého rozumu).[538]

Ďalším problémom, ktorý vyjadruje bariéru chápania zmyslu, je kreativita. V predslove tohto diela sme zo spomienok prof. Melanie Mitchell uvádzali veľké znepokojenie Douglasa Hofstadtera po skúsenosti s činnosťou programu EMI (Experiments in Musical Intelligence). EMI generovala hudbu v štýle rôznych klasických skladateľov, využívajúc veľkú množinu pravidiel, ktoré boli zamerané na zachytenie všeobecnej syntaxe i jednotlivých špecifík komponovania. Tieto pravidlá boli následne aplikované na mnohé príklady z konkrétnych diel skladateľov s cieľom vytvoriť nové diela v konkrétnom štýle konkrétneho skladateľa. A niektoré napodobeniny boli tak dobré, že oklamali aj hudobných odborníkov, z čoho pramenilo aj spomínané Hofstadterovo znepokojenie. EMI ako vysoko sofistikovaný generátor skladieb vytvorila mnoho rôznych hudobných diel, z ktorých následne autor programu, David Cope, vyberal tie najlepšie – tie, ktoré boli schopné presvedčiť aj zdatné publikum o kráse programovo vytvorenej hudby.

Môžeme povedať, že EMI bola kreatívna? Myslíme si, že v žiadnom prípade. EMI totižto netvorila originálne motívy, len sofistikovane skladala hudbu z toho, čo mala analyzované z tvorby ľudských skladateľov. EMI skladbu, ktorú vytvorila, nevedela posúdiť – z jej „pohľadu“ boli všetky rovnako kvalitne vygenerované – až Cope, resp. iný ľudský poslucháč vedeli posúdiť a vyberať tie skladby, ktoré boli dobré. EMI jednoducho nebola

537 Por. CLARK, A. Being There: *Putting Brain, Body, and World Together Again*. Cambridge, Mass.: MIT Press, 1996.

538 Takto argumentuje napr. prof. Hiroshi Ishiguro, riaditeľ Laboratória inteligentnej robotiky, ktoré je súčasťou Katedry inovácie systémov na Graduate School of Engineering Science na univerzite v Osake v Japonsku.
CHILD, A. *Origin of the Species*. [filmový dokument]. Apple TV, 2021, 5:54. [cit. 25. apríla 2022]. Dostupné na internete: <https://tv.apple.com/us/movie/origin-of-the-species/umc.cmc.1vsh8or9ojg824l4kn2kidkuw>

schopná pochopiť hudbu, ktorú generovala, a to ani v rovine hudobných konceptov, ani v rovine emočného dopadu na poslucháča.[539]

V čase písania tejto kapitoly (apríl 2022) iniciatíva OpenAI predstavila DALL-E 2[540] – systém umelej inteligencie, ktorý dokáže vytvoriť realistické obrazy a umenie na základe opisu v prirodzenom jazyku. I keď je medzi EMI a DALL-E 2 približne tridsaťročný rozdiel, v schopnosti umelej inteligencie byť kreatívnou sa vôbec nič nezmenilo. Akokoľvek zaujímavé obrazy DALL-E 2 tvorí, vôbec nechápe, čo a ako vytvára, a už vôbec nevie zhodnotiť umelecký a emočný dopad výsledku svojho algoritmu hlbokého učenia.

Na dosiahnutie kreativity nestačí len tvoriť niečo, čo môže byť označené ako kreatívne. Treba to aj vedieť pochopiť a ohodnotiť – či už na strane umelca alebo publika. Nič také však súčasné systémy umelej inteligencie nedokážu.

To isté môžeme povedať aj o rôznych aspektoch estetiky, ktoré bez prekročenia bariéry zmyslu sú technológiám AI nedostupné.

Zastávame názor, že ľudskému mysleniu sa približujúca schopnosť abstrakcie, analógie, konceptualizácie, simulácie, chápania zmyslu atď. sa nedá vytvoriť u subsymbolických systémov na základe vylepšovania súčasných modelov hlbokého učenia a masívneho prísunu trénovacích dát, resp. u symbolických systémov rozsiahlymi množinami symbolov a budovaním logických väzieb medzi nimi. Pre dosiahnutie tohoto cieľa by bolo treba prelomového vylepšenia algoritmov AI, ktoré – ako sa domnievame – súvisí so schopnosťou nielen analyticky uchopiť, ale i v reálnom svete zažiť a pochopiť niečo z podstaty myslenia a vedomia ľudskej bytosti.[541] [542]

I napriek tomu, že sa odborná komunita už celé dekády snaží, nutné podmienky

539 Por. MITCHELL, *Artificial Intelligence*, s. 274.

540 *DALL-E 2*. [on-line]. [cit. 8. apríla 2022].
Dostupné na internete: <https://openai.com/dall-e-2/>

541 Ako ANI bola výzvou pochopiť, ako tieto systémy fungujú, aký je dosah ich činnosti, aké dôsledky môžu mať na človeka a spoločnosť a ako sa k nim postaviť, aby boli eticky využívané, AGI je výzvou objaviť, ako by vlastne mala byť silná a všeobecná AI vytvorená, aký potenciál v sebe bude obnášať a čoho by bola schopná. Do akej miery môže byť všeobecná a či je vôbec reálne také niečo vytvoriť.

542 Samotný John McCarthy, jeden z hlavných protagonistov Dartmouthského seminára, už v roku 1977 uvádza, že bez veľkých **koncepčných prelomov** sa k AGI nedostaneme.
Por. SHENKER I. *Brainy robots in our future, experts think*. Detroit Free Press, September 30, 1977.

pre prekročenie bariéry chápania zmyslu, na ktoré sme doteraz poukázali, sa javia ako hlboko nepolapiteľné. Viac a viac sa preto ozývajú hlasy, ktoré prekročenie tejto bariéry spájajú s požiadavkou vedomia a sebauvedomenia, emócií, hypotézou stelesnenia, prípadne ďalšími aspektami, ktoré smerujú k analógii ľudskej osoby. A je pritom možné, že prekročiť túto bariéru znamená prijať aj niečo z limitov, emócií, logickej iracionality, socializácie a mnohých kognitívnych nedostatkov, ktoré sprevádzajú človeka a civilizáciu ako takú.[543]

Ešte nepolapiteľnejšie sa v tomto kontexte javia také méty, ako vnútorná sloboda, či zmysel pre dobro, krásu, obetu, utrpenie alebo lásku. Všetky sú navzájom prepojené a nielen z pohľadu kresťanského svetonázoru dostávajú vyšší zmysel a naplnenie v duchovnom rozmere ľudského bytia.[544]

5.2. Ontologické otázky

Organizácia Non-human Rights Project (NhRP), ktorá sa v roku 2012 pretransformovala z pôvodného Center for the Expansion of Fundamental Rights, aktívne presadzuje myšlienku začlenenia zvierat do kategórie osoby a priradenia im tomu zodpovedajúcich práv. V jednej zo svojich ostatných iniciatív spojilo sedemnásť vedcov svoje sily a adresovalo najvyššiemu súdu v New Yorku svoj pohľad (amicus brief) v mene možnosti priznania ľudských práv neľudským osobám. NhRP netvrdí, že i rôzne druhy zvierat (napr. primáty) sú ľudskými bytosťami, ale že aj ony sú osobami, ktoré majú nárok na svoje „ľudské“ práva.[545]

---

543 V podobnom duchu diskutuje aj prof. Melanie Mitchell na margo otázky, ako ďaleko sa nachádzame od vytvorenia všeobecnej umelej inteligencie na úrovni človeka.
Por. MITCHELL, *Artificial Intelligence*, s. 275-277.

544 Dokázala by umelá inteligencia trpieť? A ak áno, vedela by pochopiť a naplniť slová rakúskeho neurológa a psychiatra Viktora E. Frankla, ktorý hovorí:
„To, pred čím má utrpenie človeka chrániť, je apatia, umŕtvujúca duševná strnulosť. Dokiaľ trpíme, zostávame duševne nažive. V utrpení dokonca vyzrievame, v utrpení rastieme, činí nás bohatšími a mocnejšími.“
„V niektorých ohľadoch utrpenie prestáva byť utrpením v okamihu, keď nájde zmysel, napríklad zmysel obety.“
Viktor Frankl citáty. [on-line]. [cit. 25. apríla 2022].
Dostupné na internete: <https://citaty-slavnych.sk/autori/viktor-frankl/>

545 Por. *The Nonhuman Rights Project: Frequently Asked Questions*. [on-line]. [cit. 10. apríla 2022].

Podľa zástancov NhRP je tak pojem osoby len nádobou, do ktorej sa v okamihu, keď sa príslušnej entite prizná hodnosť osoby, môžu „nakvapkať“ príslušné práva. Legálne priznanie statusu osoby sa tak stáva základným a nevyhnutným predpokladom pre nárokovanie si akýchkoľvek práv a spoločenských možností súvisiacich s osobou. A spôsob, ako dosiahnuť tento právny status pre zvieratá, ktoré nie sú ľuďmi, je vedecky dokázať, že majú sebauvedomenie.[546]

V kontexte priznania štatútu osoby na základe sebauvedomenia sa diskusia o obsahu pojmu osoby transformuje na diskusiu o podstate sebauvedomenia.

V podaní NhRP tak ide o snahu modifikovať chápanie sebauvedomenia, aby zahŕňalo aj psychologické pochody primátov. Snahou NhRP je dokázať, že napr. spomínané primáty sú vnímavé, pričom vnímavosť (sentience) je schopnosť vnímať, cítiť a prežívať. Táto vlastnosť alebo schopnosť subjektu sa považuje za základ jeho autonómie, odkiaľ by mal byť už len malý krôčik k právnemu uznaniu statusu osoby pre primáty.[547]

Skúsme uvedené postrehy aplikovať na oblasť umelej inteligencie. Interakcia s viacerými pokročilými systémami AI by nás mohla viesť k záveru, že komunikujeme s niečím, čo dokáže byť vnímavé, dokáže reagovať na emócie a ich aj vzbudzovať, prípadne dokáže samostatne myslieť. Nemohli by sme sa tak pripojiť k argumentácii NhRP a snažiť sa pokročilé systémy umelej inteligencie[548] vyhlásiť za neľudské osoby?

Ak by sme aj hypoteticky prijali rozšírený pohľad na osobu definujúc sebauvedomenie na základe vnímavosti, dôrazne sme v predchádzajúcich kapitolách upozorňovali, že v prípade technológií AI ide vždy len o simuláciu procesov myslenia, teda aj o simuláciu čohokoľvek, čo by sme mohli nazvať súčasťou vnímavosti. Preto **zastávame názor, že ani v rozšírenom pohľade na osobu, definujúc sebauvedomenie na základe vnímavosti,**

Dostupné na internete: <https://www.nonhumanrights.org/frequently-asked-questions/>

546 THURZO, *The Influence of Existentialism and Subjectivism on the Concept of the Human Person,* s. 7.

547 THURZO, *The Influence of Existentialism and Subjectivism on the Concept of the Human Person,* s. 8. Neľudským osobám by boli spočiatku pridelené menšie práva, podľa NhRP by išlo predovšetkým o právo na fyzickú slobodu. Žiada sa nám však dodať, že sloboda ide ruka v ruke so zodpovednosťou, a tak uznanie všeobecného práva na slobodu v kontexte osoby so sebou prináša aj právnu zodpovednosť za svoje skutky. Ako však v tomto prípade realizovať penalizáciu, nápravu škody a trest, osobitne s jeho nápravným účinkom?

548 I napriek snahám niektorých odborníkov zaradiť najnovšie verzie Aplha Zero do kategórie AGI, stále sa pohybujeme v intenciách ANI.

súčasným najpokročilejším systémom ANI nemôžeme priradiť štatút osoby.

Radi by sme však podotkli, že je rozdiel medzi vnímavosťou a vedomím, resp. sebauvedomením. Vnímavosť, či senzibilita zahŕňa schopnosť subjektívneho vnímania, pocitov a skúseností. **Avšak pri vedomí, resp. sebauvedomení ide o uvedomovanie si seba a svojho okolia v plnosti kognitívnych schopností, emócií, abstrahovaní a prenášaní skúseností, chápania zmyslu, procesov učenia a rozhodovania sa, vôľových aspektov konania a pod**.

Napriek veľkým pokrokom v interdisciplinárnom rámci kognitívnej vedy stále nemáme vyriešené základné otázky, napr. vzťah medzi mysľou a telom, nevieme presne definovať schopnosť vedomia a už vôbec nie, ako ju vytvoriť na báze čisto fyzikálnych (biofyzikálnych) procesov – a to ani na úrovni vnímavosti. Takže ani nevieme, ako ju preniesť do elektronických systémov.

Mnohé aspekty ľudskej mysle sa snažíme popísať tzv. teóriou mysle a vytvoriť tak model správania, ktorého základom je zdravý rozum ako predpoklad pre všeobecnú inteligenciu. Ani takto však **nemáme odpoveď na otázku, čo je podstatou vedomia, takže nedokážeme ani určiť, aké kritériá by museli spĺňať systémy umelej inteligencie, aby sme ich mohli považovať za vnímavé a vedomé**.[549] Naviac, akýkoľvek morálny a právny systém je založený na existencii a koexistencii ľudských osôb, ktoré konajú na základe svojho svedomia, ako najbližšej normy mravnosti formovanej vonkajšou, objektivizujúcou normou, resp. normatívnym zákonom. V kontexte problémov, s ktorými sa – riešiac métu uvedomelej umelej inteligencie – potýkame, si ani nevieme predstaviť formu, resp. implementáciu mechanizmu, ktorý by etické princípy a – ak hypoteticky hovoríme o osobe – morálne zásady dokázal v rámci systému AGI realizovať.

V rámci nášho vzhľadu do ontologickej perspektívy potencionálnych uvedomelých systémov AI, by sme radi akcentovali náš náhľad, ktorý považujeme za legitímny nielen v kontexte kresťanského svetonázoru, ale i z pohľadu súčasného vedeckého bádania v oblasti umelej inteligencie.

Ľudská bytosť v jednote tela a duše je z pohľadu kresťanskej antropológie a Zjavenia stvorená na Boží obraz.[550] Ide o výnimočnosť, ktorá nás odlišuje

549 Por. THURZO, *The Influence of Existentialism and Subjectivism on the Concept of the Human Person,* s. 9.

550 *Katechizmus Katolíckej cirkvi. 362-365*. [on-line]. [cit. 10. apríla 2022].

od ostatných bytostí stvoreného fyzického sveta. Sebauvedomenie spojené s rozumom a slobodnou vôľou chápeme ako mohutnosti duše, ktoré presahujú biologickú realitu mozgu a nervovej sústavy. Na základe tejto výnimočnosti definujeme aj osobu len a výlučne ako ľudskú bytosť.

Materialistický uhol pohľadu – odmietajúc duchovnú dušu ako „formu“ tela, zásadne sa podieľajúcu na vedomí a sebauvedomení – však svojim protagonistom dáva nádej, že skôr či neskôr budeme schopní vytvoriť uvedomelú umelú inteligenciu s vedomím a sebauvedomením analogickým človeku, ktorá bude ašpirovať na štatút osoby.

Táto nádej je posilňovaná aj aktuálnymi úspechmi pokročilých systémov AI s algoritmami učenia formou odmeňovania,[551] ktoré v integrácii s ďalšími pokročilými metódami (nielen) hlbokého učenia dokážu simulovať evolúciu, resp. rekurzívne sa vylepšovať a samostatne sa vyvíjať až do podoby v súčasnosti najsofistikovanejších systémov AI.[552]

Jedným z príkladov tohto prístupu je predpoklad, že pri voľbe správnej (v súčasnosti ešte neznámej) funkcie užitočnosti (utility function) a prostredníctvom učenia vo veľmi komplexnom prostredí (na úrovni reálneho sveta) by sa systém AI mal vyvíjať smerom k dosiahnutiu kognitívnych schopností. Maximalizácia funkcie užitočnosti by následne mala znamenať dosiahnutie týchto kognitívnych schopností (videnie, rozpoznávanie, porozumenie,...). Výsledkom by bola nie komplexná AGI, ale celý komplex narrow AI algoritmov, ktoré budú oveľa lepšie ako človek.[553]

Uvedený model by sme si však dovolili rozporovať, nakoľko zastávame názor, že akýkoľvek komplex slabých AI, aj keď v danej oblasti lepších ako človek, neznamená, že ide o všeobecnú a silnú umelú inteligenciu. Veď u človeka ide o komplexnosť a integritu jednotlivých aspektov inteligencie, chápania zmyslu, atď., nie o komplex či súhrn inteligentných vlastností. **Súhrn inteligentných vlastností bez dosiahnutia koncepčných prelomov nemá potenciál prekročiť bariéru chápania zmyslu a stále**

Dostupné na internete: <https://katechizmus.sk/>

551 Učenie formou odmeňovania (učenie s posilnením, reinforcement learning) sme uvádzali v kapitole 1.6.3. a na striktne vymedzené podmienky - typy charakteristík ovplyvňujúcich činnosť inteligentného agenta, pri ktorých sú tieto systémy tak úspešné, sme upozorňovali v kapitole 5.1.

552 Najznámejším príkladom je umelá inteligencia AlphaGo Zero.

553 ROMPORTL, J. *Umělá inteligence, Life 3.0, superinteligence a život ve vesmíru*. [prednáška]. YouTube, 2019, 46:00. [cit. 7. júla 2022].
Dostupné na internete: <https://www.youtube.com/watch?v=PaLrrczJ5VI>

bude vykazovať trhliny v inteligencii, ktoré sme uvádzali v kapitole 5.1., aj keď možno v hlbšej a tým pravdepodobne aj v nebezpečnejšej rovine.

Naviac sa môžeme hypoteticky pýtať: mohol by mať takýto systém AI vedomie? Funkcionalistický prístup, ktorý u protagonistov tohoto riešenia prevažuje, to nezaujíma, keďže podľa nich sú dôležité len technické schopnosti systému.[554] Vedomie síce predpokladajú, avšak len na základe prezumpcie: čo je inteligentné, to má nejaké vedomie. Je možné však v tomto prípade hovoriť ešte o AGI? O umelej inteligencii, ktorá má byť schopná dosahovať ciele, ktoré si sama stanoví? O inteligencii na úrovni človeka, u ktorej predpokladáme, že jej podstatné aspekty sú viazané na vedomie a sebauvedomenie?

Ak hypoteticky tento pohľad rozvíjame ďalej, miesto implementácie etických princípov a požiadaviek v systémoch ANI sa nevyhneme otázke analógie svedomia a morálnych hodnôt u „možno“ uvedomelej AGI.

Ako sme uvádzali v kapitole 3.1., nepohneme sa ďalej, pokiaľ nebudeme mať vyriešený model správania sa AGI na spôsob teórie mysle a schopnosti zdravého rozumu u človeka. Narážame pritom na nepochopenie inakosti „inteligencie“ týchto technológií. Uvedomujúc si bytostné rozdiely medzi človekom a systémami AI, nemôžeme sa stavať k týmto systémom ako k niečomu ľudsky inteligentnému a človeku podobnému.

Hypotetický hodnotový rámec, morálny úsudok a mechanizmus svedomia AGI predpokladá u týchto systémov podobný model správania, ako je ten ľudský a tým aj možnosť nahradiť ľudský morálny úsudok a svedomie algoritmickým spracovaním. Pokiaľ však nebudeme mať problematiku modelu správania sa AGI (na spôsob teórie mysle a schopnosti zdravého rozumu u človeka) vyriešenú – ak by sme aj nejaký mechanizmus svedomia dokázali v AGI realizovať (resp. sa dokázal vyvinúť) nikto nám nedá záruky, že bude zdieľať rovnaké hodnoty a rovnakým spôsobom ich aj chrániť a zachovávať.[555]

554 ROMPORTL, J. *Umělá inteligence, Life 3.0, superinteligence a život ve vesmíru*, 48:00. [cit. 7. júla 2022].
Dostupné na internete: <https://www.youtube.com/watch?v=PaLrrczJ5VI>

555 Znie to síce ako z lacného sci-fi, no ako veľmi zjednodušený príklad môžeme vziať napr. rozhodnutie AGI pre pomyselnú ochranu budúcnosti ľudstva eliminovať časť obyvateľstva, ktorú by podľa svojich kritérií tento systém AGI pokladal za ohrozenie a pod. Za chápaním bizarnosti tejto úvahy stojí v mysli človeka celý hodnotový aparát stavajúci na zdravom rozume. A ak by zdravý rozum zlyhával, nastupuje

A ako sme naznačili vyššie, svedomie – ak má mať zmysel a plniť svoju funkciu – by malo mať referenčný bod mimo seba. V pohľade kresťanskej antropológie je svedomie teonómne, jeho referenčnou autoritou je Boh v rovine prirodzeného i pozitívneho zákona. **V súčasnosti nepoznáme spôsob, ako by bolo možné u umelej inteligencie dosiahnuť niečo podobné svedomiu.**

5.3. Šťastie praje pripraveným[556]

V úvode tohto diela sme spomenuli disproporciu medzi vnímaním ľudského bytia a umelou inteligenciou. Ovocím je nielen preceňovanie schopností AI, ale i neuvedomovanie si komplexnosti ľudskej inteligencie presahujúcej horizonty algoritmov a technológií.

Treba preto mať na pamäti, že dnešné pokročilé systémy umelej inteligencie sú veľmi vzdialené od všeobecnej a silnej AI, pričom o dosiahnutí schopností uvedomelej inteligencie, resp. superinteligencie môžeme len snívať.

Rizikom sa tak v súčasnosti stáva nie nebezpečne rozvinutá umelá inteligencia, ale skôr naša snaha vnímať súčasné systémy ako inteligentnejšie, než v skutočnosti sú – takmer podvedomá snaha prisudzovať im ľudské schopnosti, tieto systémy určitým spôsobom *antropomorfizovať* a následne preceňovať. Ovocím tohoto prístupu môže byť strata našej opatrnosti pri ich návrhu, vývoji, implementácii a využívaní bez plného uvedomenia si ich limitov, rizík a zraniteľností, ktoré sme v tomto diele obšírne uvádzali.[557]

---

legislatívne i praktické obmedzenie možností konkrétnej ľudskej osoby, brániace také niečo vykonať. Ak by išlo o AGI, ktorá by mala napr. na starosti armádne systémy alebo kritickú infraštruktúru, tieto obmedzenia nemusia fungovať.

556 *Louis Pasteur citáty*. [on-line]. [cit. 7. júla 2022].
Dostupné na internete: <https://citaty-slavnych.sk/autori/louis-pasteur/>

557 Por. MITCHELL, Artificial Intelligence, s. 278.
V tomto kontexte Sendhil Mullainathan, profesor ekonómie na Harvarde, uvádza:
„Mali by sme sa báť. Nie inteligentných strojov. Ale strojov, ktoré robia rozhodnutia, na ktoré nemajú dostatok inteligencie. Omnoho viac sa bojím hlúposti strojov ako ich inteligencie. Hlúposť strojov vytvára riziko chvosta (long-tail effect, ktorý sme uvádzali v kap. 2.1.2). Stroje môžu urobiť mnoho dobrých rozhodnutí a potom jedného dňa veľkolepým spôsobom zlyhajú pri nejakej udalosti, ktorá sa neobjavila v ich tréningových údajoch. To je rozdiel medzi špecifickou a všeobecnou inteligenciou. Obávam sa toho najmä preto, že sa zdá, že čoraz viac zamieňame brilantnú špecifickú inteligenciu, ktorú stroje preukazujú, so všeobecnou inteligenciou.“
The Myth Of AI: A Conversation With Jaron Lanier. [on-line]. [cit. 7. júla 2022].

Z doteraz uvedeného jasne vyplýva, že super inteligentné systémy umelej inteligencie s vlastným vedomím, resp. sebauvedomením nielenže nie sú na programe dňa, ba v kontexte kresťanského svetonázoru tvrdíme, že ani na horizonte dejín.[558]

To však neznamená, že by sme nemohli uvažovať o ďalšom rozvoji systémov umelej inteligencie smerujúc k niektorým aspektom všeobecnej a silnej umelej inteligencie. Myslíme tým predovšetkým potencionálny pokrok v oblastiach, ktoré sme uviedli na margo bariéry chápania zmyslu a koncepčných prelomov, napr. pokrok v rozvíjaní schopností abstrakcie, analógie, konceptualizácie, simulácie fungovania sveta a pod.

I keď sme sa v závere kapitoly 5.1. jasne vyhranili voči možnosti vytvoriť systémy AI schopné myslieť ako ľudia, t.j. v celej komplexnosti presahujúcej do roviny vedomia a sebauvedomenia, nebolo by možné dospieť k technológiám, o ktorých predsa len budeme môcť povedať, že myslenie už nesimulujú, ale určitým spôsobom skutočne myslia? K technológiám, ktoré by sme mohli pracovne nazvať „nevedomá“ alebo „limitovaná“ všeobecná a silná umelá inteligencia?

Ak áno, súhlasíme s prof. Stuartom Russellom, že je treba sa pripraviť na riziká s touto eventualitou spojené.[559]

Uvedomujeme si, že **kým sa dostaneme i k tzv. limitovanej AGI, musí prísť k viacerým koncepčným zlomom v oblasti AI a prelomovým posunom, resp. vedeckým objavom vo viacerých vedných odboroch**. Táto skutočnosť vedie niektorých bádateľov v oblasti vývoja pokročilých systémov umelej inteligencie k istému druhu pochybovania a skepticizmu, ktorý zachádza až k popieraniu možnosti úspechu pri dosahovaní dlhodobých cieľov, primárne k pochybnostiam o možnosti dosiahnuť aj limitovanú AGI.[560]

Dejiny vedy a pokroku poznajú veľa prípadov, keď prelomový objav, resp. jeho realizácia boli ovocím viacerých desaťročí intenzívneho výskumu a enormného finančného,

---

Dostupné na internete: <https://www.edge.org/conversation/jaron_lanier-the-myth-of-ai>

558 Ako uvádza prof. Melanie Mitchell: „Väčšina z aspektov, ktoré si na našej ľudskosti najviac ceníme, sa nedá porovnávať 's vrecom (počítačových) trikov'.“
Por. MITCHELL, Artificial Intelligence, s. 279.

559 I keď treba povedať, že prof. Russell sa v svojich úvahách nezastavuje len pri limitovanej AGI, ale snaží sa riešiť aj riziká, ktoré by mohla priniesť plne uvedomelá všeobecná umelá inteligencia.
Por. RUSSELL, *Human Compatible*, s. X.

560 Por. RUSSELL, *Human Compatible*, s. 8.

materiálneho i ľudského nasadenia. Rovnako však sú nám známe prípady, keď k prelomovému objavu prišlo v priebehu niekoľkých hodín od jasnej formulácie, resp. zverejnenia problému, prípadne prizvania ďalších riešiteľov do tímu.[561]

Aj v oblasti vývoja pokročilých systémov AI je preto nerozumné vsádzať proti ľudskej vynaliezavosti a spoliehať sa na prípadný neúspech – osobitne, ak ide o dosahovanie takej méty, akou je limitovaná AGI. Ide totiž o vývoj a tvorbu systémov, ktoré môžu byť v reálnom nasadení mocnejšie než my ľudia a prekonávať akékoľvek naše súčasné spoločenské mechanizmy.[562]

Ak by na ceste k limitovanej AGI bolo treba dosiahnuť len jeden koncepčný prelom, môže to nastať prakticky kedykoľvek a nájsť nás tak nepripravených, keďže akúkoľvek AGI s nejakým stupňom autonómnosti by sme v súčasnosti neboli schopní ani riadiť, ani ovládať.[563] **Koncepčných prelomov na ceste k AGI je našťastie viac, a preto si myslíme, že v súčasnosti nie sme v tak veľkom časovom strese, že by sme sa nemohli na limitovanú AGI komplexne a dôsledne pripravovať**.

Profesor Stuart Russell, venujúci sa i tejto oblasti vývoja pokročilých systémov umelej inteligencie, sumarizuje nasledovné koncepčné prelomy, ktoré nás čakajú na ceste k AGI:

- jazyk a zdravý rozum
- kumulatívne učenie sa konceptov a teórií
- objavovanie a spravovanie budúcich činností
- manažovanie mentálnej aktivity

561 Čítankovým príkladom je vyriešenie problému štiepenia jadra (neutrónmi indukovanej reťazovej jadrovej reakcie). V roku 1933 vyriešil výborný maďarský fyzik Leó Szilárd tento problém za menej než dvadsaťštyri hodín od chvíle, ako si pri raňajkách v dennej tlači prečítal vyjadrenie špičkového jadrového fyzika Ernesta Rutherforda o nemožnosti vyriešenia tejto úlohy (podľa Rutheforda nebolo možné potrebný proces na Zemi vytvoriť).
RHODES, R. *The Making of the Atomic Bomb*. Simon & Schuster, 1987.

562 Ako sme poukázali v kapitole 2.5., už súčasné technológie ANI v rámci sociálnych sietí dokážu premáhať ľudskú slabosť s tak negatívnymi dôsledkami, že ich reálne pociťuje celá spoločnosť.

563 Por. RUSSELL, *Human Compatible*, s. 78.

5.3.1. Koncepčné prelomy na ceste k AGI[564]

Jazyk a zdravý rozum

O inteligencii nemôžeme hovoriť bez poznania. Poznanie bytostne súvisí s jazykom, keďže poznanie sa primárne jazykom nielen prenáša, ale jazyk ho i obsahuje. Koncepčne prelomové systémy AI by preto museli vedieť nielen texty analyzovať, ale i pochopiť ich obsah, zmysel a kontext. V zásade ide o schopnosti súvisiace so zdravým rozumom, ktoré u súčasných sofistikovaných systémov umelej inteligencie vykazujú podstatné „trhliny v inteligencii“, ako sme uvádzali v kapitole 5.1.

Ak by k tomuto koncepčnému prelomu prišlo – vzhľadom na objem digitalizovaných dát a možnosti ich strojového spracovania – by technológie umelej inteligencie skutočne rozumejúce ľudskej reči dokázali v enormne krátkom čase obsiahnuť takmer všetko ľudské poznanie. Bola by to oveľa schodnejšia cesta, než všetko toto poznanie senzoricky skúmať a následne analyticky, deduktívne či synteticky objavovať.

Kumulatívne učenie sa konceptov a teórií

Ľudstvo, kráčajúc generáciami naprieč storočiami, na základe pozorovaní a skúmania vytvára koncepty a teórie, koná objavy a rozširuje tak svoje poznanie. Ide o intelektuálne procesy, ktoré sú vrstvené – koncepty a ostatné výdobytky vedy a rozvoja sú kumulované, na seba nadväzujúce a na sebe stavajúce. Ľudský mozog dokáže pre konkrétne činnosti tieto koncepty adaptívne a selektívne využívať, takže vie aplikovať koncepty adekvátne danej činnosti.

Pre sofistikované systémy umelej inteligencie, ktoré by boli schopné vo všeobecnosti riešiť problémy na intelektuálnej úrovni človeka, určite nebude stačiť súčasná architektúra čiernej skrinky, ktorá zo vstupných dát vytvorí zložité výstupné koncepty, poznanie alebo riešenia.

Pre riešenie intelektuálnych úloh na úrovni človeka budú koncepčne prelomové systémy AI musieť byť schopné kumulovať – na sebe stavať – jednotlivé stupne poznania, pričom každý z ďalších stupňov by mal byť tvorený kombináciou poznania z predchádzajúceho stupňa a spracúvania aktuálneho pozorovania (vstupov). Výsledné poznanie – koncept, teória, objav, atď. – je možné využiť nielen pre konkrétnu činnosť, ale aj ako východisko

564 Por. RUSSELL, *Human Compatible*, s. 78-93.

pre ďalší stupeň učenia sa inteligentného systému.

A naviac, koncepčne prelomové systémy, AI schopné kumulatívneho učenia sa, by na základe adaptívneho a selektívneho využívania akumulovaného poznania mali disponovať prediktívnymi schopnosťami umožňujúcimi im inteligentne konať v reálnom svete.

Nech zoberieme do úvahy akýkoľvek typ súčasných sofistikovaných systémov umelej inteligencie (symbolických alebo subsymbolických), nie sme ešte schopní realizovať kumulatívne učenie sa konceptov a teórií. Predovšetkým nemáme vyriešené tri problémy:

- aké oblasti konceptov majú byť v danej intelektuálnej činnosti zahrnuté
- ako majú byť tieto koncepty kumulované, na seba nadväzujúce a na sebe stavajúce
- algoritmicky nie sme schopní uchopiť intuíciu, vhľad a inšpiráciu, ktoré vnímame ako ľudské fenomény v inteligentnej činnosti výrazne sa podieľajúce na riešení niektorých intelektuálnych úloh stavajúcich na kumulatívnom učení[565]

Nevyriešené problémy sú podľa náročnosti uvedené vo vzostupnom poradí. Ak by sa nám podarilo vyriešiť aspoň prvé dva, znamenalo by to výrazný pokrok na ceste k limitovanej AGI schopnej určitého spôsobu myslenia.

Objavovanie a spravovanie budúcich činností

Inteligentné správanie sa v dlhom časovom horizonte nezaobíde bez schopnosti plánovať a manažovať aktivity hierarchicky vo viacnásobných úrovniach abstrakcie. Napríklad zámer vykonať cestu do Karibiku v sebe obsahuje stovky akcií na rôznych úrovniach abstrakcie, či už podľa povahy akcie (napr. zabezpečenie letenky je abstraktná aktivita pozostávajúca z mnohých konkrétnych krokov) alebo podľa časového horizontu (aktuálny nákup letenky je vnímaný i realizovaný vo všetkých parciálnych krokoch, pričom budúcotýždňové balenie je ešte len abstraktný zámer).

Aktivity nás, inteligentných bytostí sú teda organizované do komplexnej hierarchickej štruktúry pozostávajúcej z desiatok úrovní abstrakcie. A čím viac aktivít vysokej úrovne

565 Mnohí vedci v oblasti AI sa voči tomu ohradzujú, argumentujúc schopnosťou systémov AI spracovávať (sic!) množstvo hypotéz, nachádzať vhodné podklady a vypĺňať medzery novými hypotézami. Zastávame názor, že prehľadávanie hypotéz, resp. analogické postupy pre nahradenie inšpirácie, vhľadu a intuície môže fungovať v jednoduchých prípadoch. V komplexných a zložitých to však takto nie je riešiteľné. Por. RUSSELL, *Human Compatible*, s. 86.

abstrakcie poznáme a si osvojíme, tým viac dokážeme plánovať a zvládať aktivity v dlhodobom meradle.

Analogicky človeku – ako by mal inteligentný agent spravovať svoju vlastnú budúcnosť? Najbližšia budúcnosť je a musí byť mimoriadne podrobná, avšak pozerajúc o kúsok ďalej, detailov je menej. A pozerajúc ešte ďalej, plány činnosti bývajú väčšie, resp. rozsiahlejšie, ale oveľa nejasnejšie, oveľa menej podrobné, ba až vágne. Ako sa agent v čase posúva, budúcnosť sa približuje a jej plány sa stávajú viac detailnejšie. Súčasne sa do vzdialenejšej budúcnosti pridávajú ďalšie, zatiaľ ešte len vágne a menej, resp. vôbec nie podrobné plány.

Ako inteligentní agenti – súčasné sofistikované systémy AI – spravujú svoju vlastnú budúcnosť? Akým spôsobom sa realizuje konštrukcia hierarchie ich abstraktných činností?

V súčasnosti u systémov umelej inteligencie sme schopní implementovať len niektoré časti skladačky manažovania vlastnej budúcnosti. Vieme vytvoriť komplexné plány v hierarchii abstraktných činností, avšak bez uvedenej časovej osi umenšovania podrobností a nárastu vnímania šírky a spektra budúcich akcií. Ďalším kúskom skladačky, ktorý v tomto procese stále chýba, je spôsob konštrukcie hierarchie abstraktných činností, teda to, čo si musí agent uvedomiť, že treba spraviť pre dosiahnutie konkrétneho abstraktného, resp. vyššieho cieľa.

Medzi koncepčné prelomy systémov umelej inteligencie preto môžeme zaradiť i objavovanie a odhaľovanie budúcich činností, t.j. schopnosť hierarchicky manažovať aktivity od podrobného rozpracovania v prítomnosti až po viac a viac schematické uchopenie smerom do budúcnosti.

Ak by raz k tomuto koncepčnému prelomu prišlo a inteligentní agenti manažovanie budúcich akcií i abstrahovanie vyšších úrovní činnosti zvládnu, v kontexte ostatných možností umelej inteligencie budú schopní pozerať ďalej než ľudia a budú tak vedieť konať na základe rozsiahlejších a detailnejších informácií. To však bude znamenať schopnosť lepšieho reálneho rozhodovania, než ako sú schopní ľudia. V akejkoľvek konfliktnej situácii by tak systémy AI mali navrch a boli by o krok pred nami.[566]

Manažovanie mentálnej aktivity

Súčasné sofistikované systémy AI sú schopné riešiť úlohy za dosť obmedzujúcich

566 Povedané terminológiou hier šach a go – hru by sme prehrali ešte skôr, než by sme vôbec hrať začali.

podmienok ovplyvňujúcich činnosť agenta v reálnom svete. Ruka v ruke s tým ide výpočtová (hardvérová i softvérová) architektúra súčasných technológií umelej inteligencie, ktorá je síce výkonná, ale je veľmi úzko zameraná a v porovnaní s ľudským intelektom a mozgovou činnosťou nielen neefektívna pri riešení úloh vyžadujúcich skutočnú inteligenciu, ale aj neschopná zovšeobecňovať a adaptovať sa na celú šírku úloh reálneho sveta.

V súčasnosti teda nevieme, ako v sofistikovaných systémoch AI spravovať komplexné a rôznorodé aktivity reálneho sveta, ako ich integrovať a vytvárať z nich výsledky a ako alokovať výpočtové zdroje pre rôzne druhy uvažovania tak, aby správne rozhodnutia boli nielen nájdené, ale i nachádzané tou najoptimálnejšou cestou.

V prípade akejkoľvek limitovanej AGI by sme však z princípu mali uvažovať o systémoch, ktoré sú schopné zvládnuť akúkoľvek intelektuálnu úlohu a majú schopnosť generalizovať, t.j. zovšeobecňovať a prenášať, či adaptovať naučené schopnosti na iné úlohy. Koncepčne prelomové systémy AI by naviac mali dokázať manažovať svoju výpočtovú aktivitu tak, že by sa zameriavali na výpočtové celky, ktoré efektívne a rýchlo zabezpečia významné zlepšenie kvality rozhodnutí a dosahovanie cieľov.

Koncepčne prelomové systémy AI by sa mohli v reálnom svete stať tvorcami špičkových rozhodnutí,[567] keďže by boli schopné objavovať nové postupy a aktivity vyššej úrovne abstrakcie a súčasne by dokázali manažovať svoju výpočtovú aktivitu tak, že by sa zameriavali na výpočtové celky, ktoré efektívne a rýchlo zabezpečia významné zlepšenie kvality rozhodnutí. Podobne ako u ľudí by uvažovanie takýchto systémov bolo kognitívne efektívne, avšak by netrpelo malou krátkodobou pamäťou a pomalým, resp. inak limitovaným biologickým „hardvérom“, ktorý závažne obmedzuje našu schopnosť pozerať ďalej do budúcnosti, spracúvať veľké množstvo nepredvídateľných udalostí a zvažovať množstvo alternatívnych plánov.

Ak by sme k sofistikovanému systému umelej inteligencie, resp. limitovanej AGI pristupovali ako k činiteľovi (agentovi), ktorý dokáže efektívne dosahovať stanovené ciele v reálnom svete, vyriešenie doteraz uvádzaných koncepčných prelomov by malo byť dostatočnou zárukou úspechu.[568]

---

567 Išlo by o špičkové rozhodnutia, z ktorých niektoré by sme nazvali geniálne, no iné by mohli byť priam hrozivé...

568 Por. RUSSELL, *Human Compatible*, s. 93.

5.3.2. Limitovaná AGI ako konkurent človeka alebo niečo viac?

Ak sa nad doteraz uvedenými koncepčnými zlomami zamyslíme, ich prekonanie by viedlo k tak veľkému kvalitatívnemu pokroku, aký by technológie umelej inteligencie mohol postaviť takmer na roveň človeka.

Pre inteligenciu človeka, multiplikáciu jeho schopností a rozvoj civilizácie je dôležitý i sociologický rámec a spolupráca medzi ľuďmi. V prípade limitovanej AGI by i táto méta mohla byť pokorená, ak si uvedomíme, že pre efektívnu činnosť sofistikovaných strojov by sme mohli využiť schému spolupráce viacerých agentov (multi-agent cooperation design), ktorá by ešte viac mohla zvýrazniť potenciál limitovanej AGI v porovnaní so schopnosťami človeka, resp. ľudského spoločenstva.[569]

Enormný potenciál limitovanej AGI, resp. kooperujúcich systémov limitovanej AGI ľahko zvádza k prehnaným očakávaniam (alebo obavám) z takmer božských schopností týchto budúcich technológií. Nesmieme však zabúdať, že akýkoľvek objekt materiálneho sveta podlieha fyzikálnym zákonom a je súčasťou procesov a dejov, ktoré sa uskutočňujú v čase a priestore. Musíme tak rátať s technologickými limitmi samotných systémov umelej inteligencie (napr. limity použitých polovodičov, optických prenosových liniek, hardvérovej a softvérovej architektúry a pod.), s obmedzeniami reálneho sveta (napr. ako rýchlo môžu byť určité poznatky z prostredia získané, ako rýchlo konkrétne deje v reálnom svete prebiehajú,...), a tiež s nutnosťou poznania a chápania zložitých systémov (napr. biologické systémy, kvantové deje, atď.), čo by vrcholilo v extrémne náročnej úlohe vytvorenia modelu a chápania človeka.[570]

Extrémna paralelizácia a kooperácia systémov limitovanej AGI by mohla umožniť uvedené limity prekonávať lepšie a rýchlejšie ako človek, resp. celé tímy odborníkov, čo však neznamená, že by sme im mali priradiť schopnosti, ktoré presahujú fyzikálne zákonitosti a zasahujú do nadprirodzena. **Triezvy pohľad nám skôr velí uvedomiť si skutočný potenciál schopností kooperujúcich limitovaných AGI, ktorý môže byť reálnym**

569 Multi-agent cooperation design prepojených AGI by mohol zahŕňať paralelné a takmer až neobmedzene škálovateľné senzorické vstupy týchto systémov, extrémne škálovateľnú schopnosť vykonávať akcie a väčší dosah v oblasti predikcie a pohľadu do budúcnosti, atď... až po schopnosť systémov AGI riešiť problémy globálne, t.j. zapojac všetky relevantné oblasti vedy, techniky a potrebné zdroje.

570 Por. KELLY, K. *The Myth of a Superhuman AI*. In: *Wired.* [on-line]. 2017, 25. apríla. [cit. 7. septembra 2022].
Dostupné na internete: <https://www.wired.com/2017/04/the-myth-of-a-superhuman-ai/>

nebezpečím v prípade zle nastavených cieľov, pre nás nejasných procesov a postupov, technologických chýb i možného zneužitia. Trpkú ochutnávku negatívnych dôsledkov sme uvádzali v kapitole 2.5, keď sme sa v rámci rizík ANI dotkli tzv. slabej singularity, t.j. bodu, v ktorom umelá inteligencia prekonáva ľudské slabosti. A keď už schopnosť technológií ANI prekonávať ľudské slabosti je veľmi ťažko zvládnuteľným problémom súčasnej informačnej spoločnosti, čo by sme dokázali spraviť voči takým výzvam a rizikám, aké by predstavovali systémy limitovanej AGI, nehovoriac o ich kooperujúcom nasadení?

5.3.3. Limitovaná AGI ako základ žiarivej budúcnosti ľudstva

Nick Bostrom v závere svojej knihy *Superinteligencia* predikuje, že úspech v oblasti umelej inteligencie prinesie „civilizačnú trajektóriu, ktorá povedie k radostnému a triumfujúcemu vesmírnemu rozvoju ľudstva“.[571]

Pre väčšinu odborníkov z oblasti umelej inteligencie platí jednoduchá úvaha: ľudská inteligencia je zodpovedná za rozvoj našej civilizácie. Ak by sme dosiahli prístup k (v mnohých smeroch) lepšej inteligencii, mohli by sme mať rozvinutejšiu a lepšiu civilizáciu.[572]

S týmto pohľadom si však dovolíme polemizovať, keďže uvedenú úvahu považujeme za priveľmi zjednodušenú. Ak predpokladáme, že prístup k lepšej inteligencii znamená prístup k limitovanej AGI prekračujúcej koncepčné prelomy, ktoré sme uviedli v kapitole 5.3.1, v tejto úvahe ide len o čiastočný pohľad, ktorému chýba chápanie hodnôt a etického rámca, čo je oveľa podstatnejším základom pre „lepšiu“ civilizáciu než len technologické zázemie s jeho paradigmatickými dôsledkami. **Ani prekonanie spomínaných koncepčných prelomov, ktoré zaceľuje „trhliny v inteligencii“ uvádzané v kapitole 5.1, nedokáže zdolať také méty, ako vnútorná sloboda, či zmysel pre dobro, krásu, obetu, utrpenie alebo lásku. Všetky sú navzájom prepojené a nielen z pohľadu kresťanského svetonázoru dostávajú vyšší zmysel a naplnenie v duchovnom rozmere ľudského bytia a sú základom skutočného rozvoja ľudského spoločenstva**.

Paradigmatické dôsledky plne zasahujúce psychiku človeka a fungovanie spoločnosti, technologický pokrok a vyriešené najpálčivejšie problémy ľudstva, systémy limitovanej

571 BOSTROM, N. *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press, 2014, s. 260. ISBN 978-0199678112.

572 Por. RUSSELL, *Human Compatible*, s. 98.

AGI, ktoré budú schopné poskytnúť svoje služby v akejkoľvek oblasti[573] – to všetko nestačí na principiálne osobnostné i spoločenské dozrievanie a civilizačný pokrok, ktorého mierou nemôže byť len technologický pokrok alebo plošné zvyšovanie životnej úrovne.

Za horizont našich predstáv zachádzajúce možnosti systémov limitovanej AGI v ich celosvetovom nasadení môžu byť bezpochyby prostriedkami a kontextom dosahovania lepšej civilizácie. Vnímajúc memento problémov s ANI, ktoré sme uvádzali v kapitole 2.5, si však uvedomujeme i veľké riziko zneužitia tak mocných prostriedkov, akými by po prelomení koncepčných zlomov boli systémy limitovanej AGI.[574]

V ľudskej spoločnosti býva častokrát motivujúcim faktorom a činiteľom skutočného pokroku a rozvoja spoločnosti práve určitý nedostatok a náročnosť spojená s prekážkami, ktorých zdolávanie umožňuje uvedomovať si skutočné hodnoty, ich miesto a dôležitosť v živote ľudí. Aproximujúc súčasný stav rozvoja spoločnosti, boli by sme vôbec ľudsky (intelektuálne i psychologicky) a civilizačne (naprieč všetkými rozmermi spoločnosti) schopní držať krok s možnosťami AGI, ovocím jej nasadenia a dôsledkami pre naše životy?

Bez rozumného uchopenia technológií limitovanej AGI, snahy pochopiť ich riziká a predchádzať im, schopnosti pripraviť sa na dôsledky ich pôsobenia a zodpovedného nasadenia v reálnom svete by sme sa mohli dostať na civilizačnú trajektóriu, ktorá nepovedie „k radostnému a triumfujúcemu vesmírnemu rozvoju ľudstva“, ale k civilizačnému zlyhaniu bytostných rozmerov.

5.4. V čom nám AGI môže prerásť cez hlavu

Pozorný čitateľ by po prečítaní predchádzajúcich kapitol určite vedel nadviazať na nebezpečenstvá slabej umelej inteligencie, ktoré sme uvádzali v 2. kapitole a dokázal by ich kreatívne rozvinúť do rozmerov hodných pokročilej umelej inteligencie, vystatujúcej sa titulom AGI. Do oblasti týchto nebezpečenstiev môžeme načrieť i v úvode tejto kapitoly, avšak s ubezpečením, že pôjde len o slabý odvar toho, čoho by skutočná umelá

573 Všeobecne použiteľné a vo všetkých oblastiach reálneho života použiteľné systémy AGI by sme podľa vzoru z oblasti informačných technológií mohli nazývať EaaS (everything as a service – všetko ako služba).

574 Aproximácia triezveho pohľadu na súčasnú scénu informačnej kriminality, dôsledkov zneužitia kybernetických prostriedkov a rizík ich nesprávneho nasadenia na oblasť umelej inteligencie nám už teraz pridáva vrásky na čele – riziko zneužitia technológií limitovanej AGI bude enormné.

inteligencia bola potencionálne schopná.

5.4.1. Sledovanie, manipulovanie a ovládanie

V kapitole 2.5 sme uvádzali nutnú podmienku pre úspešnú činnosť sofistikovaných systémov umelej inteligencie – extrémne množstvo relevantných dát, ktoré musia byť neustále zhromažďované z reálneho sveta a priamo z ľudského prostredia. Jedným z dôsledkov bezprecedentného zhromažďovania, analýzy a spracúvania týchto informácií je ďalší posun v paradigmatickej zmene informačnej spoločnosti: žiadanou komoditou prestávajú byť informácie, ale ako produkty ich spracovania sa komoditou stávajú priamo ľudia – ich psychologické profily a vzťahy, konanie a jeho ovplyvňovanie a v konečnom dôsledku i fungovanie celej spoločnosti. So schopnosťou algoritmov AI vytvárať modely, ktoré dokážu predpovedať naše reakcie a správanie prichádza i možnosť a snaha ovplyvňovať naše vnímanie, rozhodnutia a konanie.

V rámci znalostnej a informačnej spoločnosti ponorení do kybernetického sveta si ani neuvedomujeme, ako je naša myseľ a psychika zraniteľná, a tak postupne prechádzame od technologického prostredia založeného na systémoch AI k prostrediu založenému na závislosti a manipulácii. V tomto kontexte sa preto obávame tzv. „slabej“ technologickej singularity, v ktorej technológie AI ovládnu a prekonajú naše slabosti – už vtedy totiž prichádza víťazstvo umelej inteligencie a rodí sa porážka ľudstva.[575]

A nielen to. V kontexte algoritmického riadenia štátu a v neustálej snahe akcentovať bezpečnosť na úkor slobody sa v súčasnosti i v rozvinutých demokraciách pristupuje k využívania systémov AI na čoraz rozsiahlejšiu kontrolu a monitoring obyvateľstva, sledovanie technologických stôp jednotlivcov a sofistikovanú analýzu i predikciu správania obyvateľstva.[576] Fenomén umelej inteligencie sa tak podieľa nielen na civilizačnom pokroku so všetkými rizikami a výzvami, ktoré to obnáša, ale je súčasťou systémových zmien fungovania modernej civilizácie.

Toto všetko sa v súčasnosti uskutočňuje vďaka systémom slabej umelej inteligencie (ANI).

Skúsme si predstaviť, akou revolúciou v oblasti sledovania a dohľadu, analýzy a predikcie by bolo dosiahnutie koncepčných prelomov, ktoré sme uvádzali v kapitole 5.3.1.

---

575 Dva pohľady na singularitu v oblasti umelej inteligencie sme schematicky vyjadrili na obr. č. 10.

576 Algokraciu, t.j. algoritmické riadenie štátu a rôzne aspekty plošného sledovania, analýzy a predikcie ľudského správania sme rozoberali v kapitole 2.6.

Zhromažďované dáta v súčasnosti spracúvajú algoritmy ANI, ktoré nie sú schopné chápať ich obsah. Limitovaná AGI, ktorá by dokázala chápať obsah, zmysel a kontext, by nebola ďaleko od „činnosti“ zdravého rozumu a ďalekosiahlych dôsledkov, ktoré z toho vyplývajú.

Išlo by o inteligentné systémy, ktoré to, čo spracúvajú, chápu a sledujúc zadané ciele, môžu na základe toho veľmi adekvátne reagovať. Išlo by tiež o systémy, ktoré by za základe obsiahnutého poznania celej ľudskej civilizácie a kumulatívneho učenia sa konceptov a teórií nemali problém vnímať rozpor medzi monitorovanými činmi objektov (ľudí i strojov) s obmedzeným poznaním, resp. zmyslovým či senzorickým vnímaním a optimálnymi krokmi, ktoré by boli ovocím takmer plného poznania a chápania prebiehajúcej reality. A vzhľadom na zvládnuté ďalšie koncepčné prelomy, osobitne manažment budúcich činností a mentálnej aktivity, by tieto systémy dokázali byť vo svojich predikciách i reakciách efektívnejšie než ľudia a v optike zadaných cieľov by sa stávali tvorcami špičkových rozhodnutí. Problémom je však, čo by bolo zámerom, resp. cieľom limitovanej AGI a či by ovocie, ktoré by jej rozhodnutia priniesli, nebolo pre ľudskú civilizáciu príliš trpké.

5.4.2. Manipulácia a ovládanie nášho správania

Osobitne treba zdôrazniť riziko manipulácie a ovládania nášho správania, ktoré v kontexte ANI bolo v kapitole 2.5 vyjadrené v celej šírke spektra, od manipulačných algoritmov sociálnych sietí až po fenomén *deepfake* a falošné identity, a v kapitole 2.6 zavŕšené odmeňovaním i trestaním na základe tzv. sociálneho kreditu ako jednej z pokrivených foriem algokracie.

Dôsledkom zvládnutia koncepčných prelomov uvádzaných v kapitole 5.3.1 by nastal bezprecedentný kvalitatívny posun v schopnostiach umelej inteligencie manipulovať a ovládať, keďže na miesto úzko špecializovaných seba zdokonaľujúcich algoritmov by nastúpili systémy, ktoré daným zadaniam rozumejú a dokážu ich s vysokou efektivitou riešiť ako intelektuálny problém, v ktorom majú všetky výhody na svojej strane.

Mnohí protagonisti všeobecnej umelej inteligencie by jasali, že vďaka dosiahnutiu koncepčných prelomov môžu nasadiť také systémy, ktoré by mohli byť odolné voči zneužitiu na nesprávne ciele. Existuje však problém, ktorý ich zápal a nadšenie môže úplne zahasiť.

Ako sme uvádzali v kapitole 5.2, pokiaľ u limitovanej AGI nebudeme mať vyriešený

model správania analogický ľudskému a nebudeme vedieť ako do systému vtlačiť základné hodnotové rámce, absolútne sa nemôžeme na umelú inteligenciu spoľahnúť, mysliac si, že sama dokáže vyriešiť riziká, ktoré z jej implementácie do mechanizmov modernej spoločnosti vyplývajú.

Osobitne rizikovým faktorom budú parciálne, tzv. inštrumentálne ciele[577], ktoré si pokročilý systém AI stanoví na dosiahnutie zadaných úloh. Úloh, ktoré – ako predpokladáme – sa budeme snažiť čo najlepšie definovať a strojovú snahu o ich zmenu efektívne korigovať.[578]

5.4.3. Právo na psychickú bezpečnosť

Človek ako súčasť modernej informačnej spoločnosti trávi čoraz viac zo svojho života v kybernetickom priestore.[579] V tom priestore, ktorý bude čím ďalej tým viac tvorený a manažovaný systémami umelej inteligencie.

V kontexte rizík, ktoré môžu znamenať zásahy a pôsobenie limitovanej AGI do virtuálneho priestoru, prof. Stuart Russell uvádza ako jedno z veľkých nebezpečenstiev aj ohrozenie psychickej bezpečnosti a preto po vzore 3. článku Všeobecnej deklarácie ľudských práv[580] volá po práve na psychickú bezpečnosť ako práve žiť v prevažne pravdivom informačnom prostredí.[581]

Ako ľudské bytosti máme nielen bytostnú tendenciu veriť našim zmyslovým vnemom – čo vidíme a počujeme, ale na tomto zmyslovom vnímaní je postavená aj činnosť nášho intelektu v abstrahovaní, vytváraní súdov a úsudkov, konceptov, predikcií a pod.

Problémy, ktoré sme v kapitole 2.5 rozoberali, poukazujú na extrémnu zraniteľnosť ľudského intelektu v oblasti zmyslových vnemov a z toho vyplývajúce mnohé nežiadúce

---

577 Problematike inštrumentálnych cieľov sa venujeme o niekoľko strán ďalej.

578 Že to vôbec nie je jednoduché, budeme rozoberať v rámci tzv. problému kráľa Midasa.

579 Por. ŠANTAVÝ, *Analýza virtuálneho sveta v kontexte náboženskej formácie, pôsobenia a života Cirkvi* [on-line], s. 20-22. [cit. 12. septembra 2022].
Dostupné na internete: <https://peter.santavy.cloud/data/uploads/docs/analyza_virtualneho_sveta_v_kontexte_nabozenskej_formacie_posobenia_a_zivota_cirkvi.pdf>

580 Článok 3: „Každý má právo na život, slobodu a osobnú bezpečnosť.“
Všeobecná deklarácia ľudských práv. [on-line]. New York, 10.12.1948. [cit. 12. septembra 2022].
Dostupné na internete: <https://www.ustavnysud.sk/documents/10182/992240/DE01_48.pdf/>

581 Por. RUSSELL, *Human Compatible*, s. 108-110.

dôsledky pre psychiku človeka a jeho zaradenie sa v spoločnosti. Jednoducho ako ľudia sme extrémne zraniteľní technologickým rámcom manipulovania vnímanej reality.

Ako sme už v tejto kapitole uviedli, **dôsledkom zvládnutia koncepčných prelomov uvádzaných v kapitole 5.3.1 by nastal bezprecedentný kvalitatívny posun v schopnostiach umelej inteligencie manipulovať a ovládať, teda vytvárať pozmenenú realitu, ktorú by ľudský mozog nedokázal adekvátne spracúvať. Limitovaná AGI by bola schopná vytvárať virtuálnu realitu podľa svojich vlastných kritérií a v kvalite nerozoznateľnej od pravdivých informácií a skutočnej reality. Samozrejme, koncepčne prelomová limitovaná AGI by mala potenciál tieto pokrivenia aj odhaľovať, no znovu stojíme pred problémom, ako zabezpečiť, aby v snahe o dosahovanie stanovených cieľov sa nevyberala temnými cestami manipulovania kybernetického priestoru a do ľudských životov zasahujúcej tzv. rozšírenej reality, čoho ovocím by mohli byť nesmierne škody na ľudskej psychike a spoločnosti**.[582]

5.4.4. Smrtiace autonómne zbrane

V kapitolách 2.7.5 až 2.7.7 sme dosť podrobne rozoberali problematiku smrtiacich autonómnych zbraňových systémov (LAWS) a kybernetických zbraní poháňaných technológiami umelej inteligencie, keďže ich existencia, resp. možnosť nasadenia by významným spôsobom mohla ohroziť a redukovať ľudskú bezpečnosť na všetkých úrovniach – osobnej, lokálnej, národnej i medzinárodnej.

Na základe koncepčných prelomov AI sa však riziko ohrozenia bezpečnosti ľudskej civilizácie môže stať existenčným – limitovaná AGI by totiž mala potenciál integrovať, ovládnuť a zneužiť LAWS na dosahovanie svojich vlastných cieľov.

Znovu pripomíname, že základným problémom nemusia byť stanovené celkové ciele, ale ich parciálne derivácie, t.j. inštrumentálne ciele, ktoré by umelá inteligencia považovala za potrebné realizovať ako míľniky na ceste k dosahovaniu cieľov, o ktorých samozrejme prezumujeme, že by mali byť pre ľudskú civilizáciu i jednotlivcov osožné.

582 Medzi základné piliere každá zdravej a zrelej spoločnosti patrí pravda a spravodlivosť. Podkopanie týchto pilierov následne vedie k nepríjemným dôsledkom pre celú spoločnosť.

5.4.5. Eliminovanie práce tak, ako ju poznáme

Jednou z výziev, s ktorou sa nielen v rámci automatizácie a robotizácie, ale aj predtým pri každej industriálnej zmene potýkame, je ohrozenie, ktoré nové technológie znamenajú pre trh práce. V retrospektíve riešení posledných dvesto rokov by v zásade malo ísť o tému, ktorá by prinášala vrásky na čele len ekonómom a priemyselným stratégom zanedbávajúcim proporčné plánovanie rozvoja priemyslu a hospodárstva, resp. investíciu do jeho reštrukturalizácie a oblastí znalostnej spoločnosti.[583]

Avšak pri nasadení umelej inteligencie, ktorá je schopná prekročiť koncepčné prelomy, je to oveľa zložitejšie. Limitovaná AGI totižto dokáže zastať i tie nové pozície, o ktorých sme v rámci doterajšej automatizácie tvrdili, že budú pripravené pre ľudí, ktorých doterajšiu prácu nahradia stroje. Všeobecná umelá inteligencia má totiž potenciál nahradiť nielen bežné, prevažne fyzické práce, vhodné pre automatizáciu a robotizáciu, ale dokázala by eliminovať aj intelektuálne práce a viaceré služby viazané na kognitívne, intelektuálne, prípadne do určitej miery emočné aspekty ľudského mozgu. A to platí i v prípade, že by sme na týchto pozíciách rátali s kooperáciou ľudí a technológií limitovanej AGI.[584]

Z doterajších priebehov priemyselných revolúcií a zavádzania strojov badať, že do určitej miery s automatizáciou rastie i schopnosť lacnejšie a efektívnejšie vykonávať prácu dokonca s nasadením viacerých zamestnancov. Od určitej miery technologického pokroku v danej oblasti zamestnanosť začína klesať vzhľadom na rozvíjajúce sa schopnosti technológie.[585] Žiaľ, evidencia pádu zamestnanosti v ostatných štyridsiatich rokoch vo všetkých odvetviach, ktoré sa pre zvyšovanie produktivity vybrali cestou

583 Žiaľ, za jeden z príkladov môžeme považovať i súčasné Slovensko s jeho extrémnou orientáciou na montážne automobilové linky, ktoré sa v rámci rozširujúcej automatizácie a robotizácie veľmi ľahko zmenia na podniky spravované za pomoci slabej umelej inteligencie, pričom ľudský personál nebudú takmer vôbec vyžadovať.

584 Pomerne rozsiahlu analýzu a sumarizáciu týchto problémov obsahujú publikácie:
FORD, M. *Rise of the Robots: Technology and the Threat of a Jobless Future*. Basic Books, 2015.
CHACE, C. *The Economic Singularity: Artificial Intelligence and the Death of Capitalism*. Three Cs, 2016.

585 Matematickú analýzu tohoto procesu uvádza James Bessen.
BESSEN, J. *Artificial intelligence and jobs*. In: AGRAWAL, A., GANS, J., GOLDFARB, A. *The Economics of Artificial Intelligence: An Agenda*. [on-line]. 2019. [cit. 14. septembra 2022].
Dostupné na internete: <https://www.nber.org/papers/w24235>

technologického rozvoja, tento regres zamestnanosti len potvrdzuje.[586] Nemusíme pripomínať, že slovné spojenie „rozvíjajúce sa schopnosti technológie“ je v prípade nástupu limitovanej AGI len veľmi miernym eufemizmom...[587]

Psychologický aspekt práce v ľudskom živote a jej spoločenský rozmer sú tak významné a dejinne fixované, že s nahradením ľudskej činnosti technológiami limitovanej AGI a parciálnou snahou o riešenie nezamestnanosti (napríklad prostredníctvom univerzálneho základného príjmu) sa môže civilizácia dostať na pokraj disruptívnych zmien nielen v identite človeka a jeho schopnostiach realizovať ľudské bytie vo svojom rozvoji a naplnení, ale aj v celej škále problémov, ktoré môžu v spoločnosti vyvstať a bez riešenia ktorých riskujeme neudržateľnú úroveň sociálno-ekonomickej dislokácie.[588]

5.4.6. Humanoidný výzor limitovanej AGI v interakcii s človekom

V kapitole 5.1 sme ako jeden z fenoménov snahy o vytvorenie umelej inteligencie na úrovni človeka uvádzali tzv. hypotézu stelesnenia založenú na premise, že stroj nemôže dosiahnuť inteligenciu ľudskej úrovne bez nejakého druhu „tela“, v ktorom interaguje s okolitým svetom, pretože mozog oddelený od tela nikdy nemôže nadobudnúť koncepty potrebné pre všeobecnú inteligenciu.

Myslíme si, že nie je problém v navrhovaní robotických systémov AI, avšak na základe uvedenej snahy a v podvedomej túžbe po dosiahnutí umelej inteligencie, ktorá sa vyrovná, resp. je podobná človeku, u systémov AI v nemalej miere prichádza k napodobňovaniu človeka, fyzického výzoru a nonverbálnej komunikácie.

586 AUTOR, D., SALOMONS. A. *Is Automatization labor-displacing? Productivity growth, employment, and the labor share*. In: *Brookings Papers on Economic Activity*. [on-line]. Spring, 2018, s. 1-63. [cit. 14. septembra 2022].
Dostupné na internete: <https://www.jstor.org/stable/26506212>

587 Preto vo svete pribúdajú centrá, ktoré sa tomuto problému seriózne venujú, z tých najznámejších napr. Work and Intelligent Work and Systems group in Berkeley, projekt Future of Work and Workers v Center for Advanced Study in the Behavioral Sciences at Stanford, alebo Future of Work Initiative at Carnegie Mellon Univeristy.

588 Ide o celý komplex problémov zahŕňajúci ľudskú slobodu, dôsledky nečinnosti, stratu zmyslu a dosahovania cieľov, rôznorodosť ľudských schopností v porovnaní so zostávajúcimi možnosťami pracovného zaradenia, disproporcia v rozdeľovaní príjmov a pod.
Pomerne podrobne túto tému rozoberá prof. Stuart Russell.
Por. RUSSELL, *Human Compatible*, s. 113-124.

Zdieľame pohľad viacerých protagonistov vývoja umelej inteligencie, podľa ktorých sa zdá byť vytváranie systémov AI na spôsob imitácie ľudí skutočne zlým nápadom.[589]

V zásade ide o nasledovné problémy:

- **v interakcii s umelou inteligenciou imitujúcou človeka je veľké riziko prechodu z roviny rozumovej, a tým aj vedomej opatrnosti, do roviny emocionálnej**. Keď už systémy ANI dokážu v človeku evokovať emociálne prejavy a naviazanie, ovplyvňujúce nielen rozhodovanie, ale i psychické rozpoloženie, schopnosťami limitovanej AGI vybavené imitácie ľudských bytostí by mali potenciál ešte viac manipulovať a ovplyvňovať naše rozhodovanie a správanie.
- **na základe imitácie ľudských bytostí človeku sa podobajúcim systémom limitovanej AGI je vážny predpoklad prisudzovať týmto strojom ľudské schopnosti, hodnotový systém a model správania**. Už viackrát sme viaceré aspekty tohoto vážneho rizika spomínali, avšak v spojení AGI s ľudským výzorom jeho pravdepodobnosť enormne stúpa.
- **ľudský výzor a tzv. telesná humanizácia umelej inteligencie môžu zvyšovať tlak na postavenie systémov AGI na roveň človeka so snahou o udeleniu štatútu osoby**.[590] Dôsledkom môže byť delegovanie rozhodovacieho postavenia v spoločnosti, a tým aj k civilizačnému stretu – incidenty, ktorých podstatou by bolo nepochopenie a následné zlé rozhodnutia AGI atakujúce ľudskú dôstojnosť a degradujúce postavenie človeka.
- **pokročilá umelá inteligencia sa môže naučiť využívať psychologické aspekty interakcie svojho humanoidného prevedenia s človekom na dosiahnutie svojich cieľov**. Môžeme síce znovu predpokladať, že dokážeme zabezpečiť, že by stanovené celkové ciele činnosti limitovanej AGI boli pre ľudskú civilizáciu i jednotlivcov osožné, no zároveň si môžeme byť takmer istí schopnosťou AGI spoznať a chcieť využiť ovplyvňovanie svojím humanoidným výzorom na dosahovanie inštrumentálnych cieľov.[591]

Možnosť vytvárať humanoidné systémy pokročilej umelej inteligencie je síce výzva

589 Por. RUSSELL, *Human Compatible*, s. 124-131.

590 Problematike udelenia štatútu osoby systémom umelej inteligencie sme sa venovali v kapitole 5.2.

591 Tento fenomén je evidentný u niektorých algoritmov slabej AI nasadených v rámci sociálnych sietí.

ambiciózna a lákavá, no mali by sme si dvakrát rozmyslieť, či týmto spôsobom umožníme limitovanej AGI fungovať a budeme tak zvyšovať riziko manipulácie a ovplyvňovania zo strany AI a zahrávať sa s možnou degradáciou ľudskej osoby a nepredvídateľnými psychologickými i sociologickými dôsledkami pre spoločnosť.

5.4.7. Človek a gorila – vymenené úlohy

Skúsme zapojiť predstavivosť a všimnime si, že po evolučnej stránke pred desiatkami miliónov rokov sme mali s dnešnými gorilami spoločných predkov, až kým sa genetické línie nerozdelili. Ak by sme hypoteticky umožnili gorilám, aby o tom rozmýšľali a uvedomili si svoju dnešnú situáciu, ako by sa asi títo naši „spolupútnici“ vo vývoji cítili? Okrem rôznych iných konotácií by určite veľmi negatívne vnímali, že bez ľudskej blahosklonnosti v zásade nemajú žiadnu šancu na prežitie. Ak teda existuje vyššia inteligencia – a tou sme my – gorily i iné druhy prežívajú len preto, že im to umožňujeme a ešte sme ich nevyhubili.

Uvedomujeme si, že naša dominancia a vláda nad svetom je výsledkom našej inteligencie, takže vytvorenie niečoho, čo by bolo určitým spôsobom inteligentnejšie než ľudia, nemusí byť až tak dobrá idea. V tomto ponímaní sa výdobytky a benefity AGI môžu stať mrazivým komfortom pre ľudstvo, ktoré by začalo byť odkázané na milosť a nemilosť strojovej inteligencie.[592]

Už viac než 150 rokov vedú obavy z ovládnutia, prípadne eliminácie ľudstva inteligentnými strojmi k volaniu po ich zákaze, čo sa vo vývoji algoritmov pokročilej umelej inteligencie pretavuje v apel na zákaz vývoja AGI.[593]

---

592 Prof. Russell to nazýva problém goríl (gorilla problem), t.j. problémom, či ľudia dokážu udržať svoju nadvládu a autonómiu vo svete, v ktorom by boli stroje s podstatne vyššou inteligenciou ako ľudia. RUSSELL, *Human Compatible*, s. 132.

593 Existenčný „problém goríl“ má svoj základ v obavách sprevádzajúcich nasadenie samostatne pracujúcich strojov už v polovici 19. storočia – či už išlo o nejasnú obavu priekopníkov programovania Charlesa Babbage a Ady Lovelace (1842), jasnejšie kontúry rizika inteligentných strojov v článku Richarda Thorntona (1847) alebo seriózu polemiku ohľadom existenčných rizík výpočtových zariadení v novele Erewhon od Samuela Butlera (1872). V pionierskych časoch zrodu moderných počítačových systémov a kladenia základov vývoja umelej inteligencie na tieto riziká upozorňoval a seriózne obavy predkladal i jeden z najfenomenálnejších mysliteľov úsvitu počítačov a AI, Alan Turing.
TURING, A. *Computing machinery and intelligence*. Mind 59, 1950, s. 433-460.
THORTON, R. *The age of machinery*. Primitive Expounder IV, 1847, s. 281.
TURING, A. *Intelligent machinery, a heretical theory*. The 51 Society, Manchester, 1951.

Akékoľvek nadšenie pre zákaz vývoja AGI sa však veľmi rýchlo rozplynie tvárou v tvár súčasnej realite vývoja umelej inteligencie. A to z viacerých dôvodov.

Ponajprv ide o ekonomickú stránku veci. Očakávaná ekonomická hodnota pokročilej umelej inteligencie sa ráta v biliónoch eur[594], takže korporátny, politický i zbrojársky tlak na vývoj všeobecnej AI je a bude enormný. A naviac – čím ďalej sa vo vývoji a súbežnom aplikovaní technológií pokročilej AI budeme nachádzať, tým väčšie, ba až nenahraditeľné škody by vypnutie týchto systémov pre modernú spoločnosť mohlo znamenať.

Ďalším, ešte dôležitejším faktorom je skutočnosť, že vlastne ani nevieme, čo by sme mali zakázať. Progres vo vývoji a parciálne úspechy na ceste k zdolaniu koncepčných prelomov sa vyskytujú naprieč mnohými odbormi matematiky, neurovedy, základného i aplikovaného výskumu počítačových vied, vývoja systémov slabej umelej inteligencie i robotov, pri riešení úplne iných výskumných zadaní a pod. V tak širokom meradle nie sme schopní a ani nevieme čo zakázať, resp. obmedziť.

K problému goríl nám tak zostáva jediný prístup – na vývoji limitovanej AGI neustále pracovať s vedomím, že chápeme riziká spojené s všeobecnou umelou inteligenciou a že si uvedomujeme, že vytvorenie niečoho, čo by bolo určitým spôsobom inteligentnejšie než ľudia, nemusí byť až tak dobrá idea. Na základe histórie ľudskej civilizácie však máme obavy, že i napriek evidencii tisícročí vývoja, koexistencie i zápasov rôzne inteligentných biologických druhov na Zemi, budeme mať dosť veľký problém tento prístup zachovať.

5.4.8. Čo kráľ Midas nedomyslel

Podobne, ako sa obavy z gorila problému v rôznych formách vynárali už od úsvitu využívania výpočtových zariadení, pionieri počítačov a kybernetiky si uvedomovali i ďalší, **nemenej závažný problém – problém s dosahovaním stanovených cieľov a našich skutočných zámerov**.

Norbert Wiener, priekopník v oblasti umelej inteligencie, kognitívnej vedy a teórie riadenia bol už od svojej mladosti znepokojený **nepredvídateľnosťou komplexných systémov pracujúcich v reálnom svete plnom rôznorodej dynamiky a nepreberného množstva ovplyvňujúcich faktorov** a obával sa nadmernej dôvery vtedajších (a treba povedať, že i dnešných) vedcov a inžinierov v svoje schopnosti ovládať a mať pod kontrolou

594 Por. RUSSELL, *Human Compatible*, s. 75, 93, 136, 163.

komplexné technické zariadenia. Tejto téme sa venoval v sérii svojich článok, ktoré v šesťdesiatych rokoch minulého storočia vyvrcholili v názore, že **nie sme schopní presne a kompletne definovať skutočné ľudské ciele, teda ani tie, ktoré stanovujeme systémom umelej inteligencie**.[595]

Z uvedenej premisy by však vyplývalo, že štandardný model práce systémov AI, v ktorom sa pokúšame strojom stanovovať svoje vlastné ciele, je odsúdený na neúspech.[596]

Na dôvažok, pri stanovovaní cieľov pokročilým systémom limitovanej AGI máme tendenciu predpokladať u nich analógiu ľudského modelu správania, takže stanovovanie cieľov kalkuluje s ich chápaním v našom ľudskom ponímaní, z čoho vyplýva ďalší „stupeň voľnosti“ týchto systémov, t.j. ďalšia množina možných zlyhaní pri dosahovaní zadaných cieľov.

Problém s dosahovaním cieľov by sme mohli obrazne nazvať problémom kráľa Midasa. Nemyslíme tým niektorého zo starovekých panovníkov Frýgie, ale o mytologický príbeh, v ktorom si kráľ Midas ako odmenu od boha Dionýza žiadal, aby všetko, čoho sa dotkne, sa premenilo na zlato. A dostal presne to, čo žiadal, neuvedomujúc si, že na zlato premenil osoby, ktoré miloval i jedlo a nápoj, bez ktorých následne zomieral. Zámer kráľa Midasa bol jasný – stať sa najbohatším človekom na svete. Spôsob, ktorým si to žiadal dosiahnuť a ktorý aj skutočne viedol k nesmiernemu bohatstvu, bol pre kráľa i jeho okolie deštrukčný.

V kontexte Wienerových obáv z našej neschopnosti presne a kompletne definovať skutočné ľudské ciele a s určitou nepredvídateľnosťou komplexných systémov pracujúcich v reálnom svete je problém kráľa Midasa neľahkou výzvou pre vývoj akýchkoľvek kybernetických strojov a inteligentných systémov. Prekročiac koncepčné prelomy by sa totiž limitovaná AGI mohla snažiť dosahovať zadané ciele skutočne obludným spôsobom.

U systémov ANI nie sú principiálne prekážky sa s týmito rizikami – problémom kráľa Midasa – vysporiadať, či už ide o ich predikciu, odhalenie pri simuláciách a testovaní alebo možné odstránenie chýb zistených počas prevádzky, nehovoriac o tom, že ich dosah

595 Por. WIENER, N. *The Human Use of Human Beings*. Riverside Press, 1950.
Por. WIENER, N. *God and Golem, Inc.: A Comment on Certain Points Where Cybernetics Impinges on Religion*. MIT Press, 1964.

596 Por. RUSSELL, *Human Compatible*, s. 136-140.

na reálny svet je stále dosť obmedzený.

U systémov limitovanej AGI, osobitne u systémov s globálnym dosahom ich činnosti, je to iné – neboli by sme schopní ich činnosť simulovať, reálne determinovať alebo počas činnosti meniť. Väčšinou by sme ani nevedeli určiť, akou cestou optimalizácie a výberu spôsobu riešenia problémov sa umelá inteligencia pri dosahovaní stanovených cieľov vyberie.

A ak by sme chceli systémy AGI regulovať a vylúčiť nesprávne postupy a riešenia, v kontexte teórie riadenia procesov v reálnom svete by pri riešení komplexných problémov bolo veľmi ťažké, ba až nemožné, vopred predvídať a vylúčiť všetky katastrofálne spôsoby, ktoré by si stroje mohli vybrať pri dosahovaní stanovených cieľov.

Dokonca by sme mohli tento problém transponovať a zájsť až tak ďaleko, že ak by sme mali určitý konkrétny cieľ a systém AGI iný, ktorý je v konflikte s naším, stroj by si mohol presadiť svoj cieľ ako ten, ktorý je podľa neho v kontexte nášho zadania pre nás určite vhodnejší. To by sme sa však už dostávali do bodu, v ktorom by sme začali zisťovať, ako by sa asi cítili gorily...

5.4.9. Mínové pole inštrumentálnych cieľov

Viackrát sme už spomenuli riziká súvisiace s parciálnymi, resp. inštrumentálnymi cieľmi. Skúsme si posvietiť na dve riziká, ktoré by sme u umelej inteligencie asi nečakali – strach a chamtivosť.

Strachom môžeme nazvať určitú logickú dedukciu systému, ktorý „si uvedomuje“, že ak má stanovené ciele dosiahnuť, musí zostať funkčný. U systémov limitovanej AGI preto musíme seriózne predpokladať, že ak majú byť schopné vykonať zadané úlohy a dosiahnuť stanovené ciele, budú sa snažiť a častokrát budú i schopné sa chrániť voči výpadku a akémukoľvek zasahovaniu do svojej činnosti. Z toho vyplýva pre nás veľmi nepríjemné zistenie, že ak ciele nebudú správne stanovené, resp. umelou inteligenciou pochopené, ťažko bude činnosť takéhoto systému obmedziť, usmerniť alebo zastaviť.[597]

Pre takéto správanie netreba do systémov limitovanej AGI implementovať žiadne seba

597 V diaľke sa tak rozplýva riešenie, ktoré navrhol Alan Turing – možnosť vypnúť systém AI v strategických momentoch.
Por. RUSSELL, *Human Compatible*, s. 141.

chrániace a sebaobranné mechanizmy.[598] **Toto správanie sa stane logickým dôsledkom dostatočne pokročilej inteligencie, ktorá bude chcieť spraviť všetko pre to, aby splnila stanovený cieľ. Sebaochrana, sebazáchova a sebaobrana sa úplne logicky u limitovanej AGI budú stávať parciálnym, resp. inštrumentálnym cieľom a prostriedkom pre dosahovanie stanovených cieľov**.[599]

Podobne ako sebaobrana, aj snaha o zabezpečenie potrebných prostriedkov pre svoju činnosť sa logicky môže stať inštrumentálnym cieľom a nástrojom na ceste k dosahovaniu stanovených cieľov. Táto „túžba“ v snahe o dostatočné zabezpečenie činnosti sa môže z nášho pohľadu stať až neobmedzenou a nezriadenou, ľahko nazvateľnou chamtivosťou.

S každým dosiahnutým prostriedkom ako uskutočneným inštrumentálnym cieľom sa na ceste k naplneniu stanovených cieľov môžu objaviť ďalšie parciálne ciele, resp. prostriedky, ktoré treba zabezpečiť. V prípade veľkých a komplexných definitívnych cieľov môže postupnosť zabezpečovania inštrumentálnych cieľov mať expanzný charakter až do tej miery, že naplnenie týchto parciálnych cieľov môže činnosť AGI postaviť do konfliktu s ľuďmi. Chamtivosť inteligentných strojov tak môže prerásť do špirály procesov, ktoré AGI môžu postaviť do nevyhnutného konfliktu s ľudskou civilizáciou, pričom zo skôr uvedených možností a rizík AGI je jasné, kto by mal navrch.[600]

5.4.10. Evolučné analógie[601]

Skúsme sa ešte raz vrátiť k evolučným analógiám, ktoré sme využili pri opise tzv. Gorila problému v kapitole 5.4.7. Použijúc súčasné poznatky z prírodných vied, tzv. adaptívna radiácia v ringu evolučného vývoja, sprevádzaná zápasom jednotlivých druhov o čo najlepšie prispôsobenie sa meniacemu sa prostrediu, je zavŕšená a prekročená

598 Všimnime si, že Asimovov 3. zákon robotiky: „robot musí chrániť svoju vlastnú existenciu“, je u AGI úplne zbytočný.
ASIMOV, I. *Runaround*. Astounding Science Fiction, marec 1942.

599 OMOHUNDRO, S. *The basic AI drives*. In: *Artificial General Intelligence 2008: Proceedings of the First AGI Conference*. IOS Press, 2008.

600 Por. RUSSELL, *Human Compatible*, s. 141-142.

601 Prvotnou, veľmi voľnou inšpiráciou pre napísanie tejto kapitoly bola prednáška *O (s)tvorení prírody s nadhľadom alebo ako Stvoriteľ ostáva najgeniálnejším vedcom* v podaní prof. RNDr. Petra Fedora, DrSc. Prednáška odznela 20. októbra 2022 v rámci Fakultných štvrtkov na RKCMBF UK v Bratislave.

inteligenciou. **Inteligencia v zásade najlepšie uschopňuje adaptovať sa a víťaziť v evolučnom zápase**.[602]

Ak – zjednodušene povedané – už klasický evolučný boj v sebe zahŕňa konkurenčný zápas určitých inteligencií i súťaž schopností prispôsobiť sa meniacim sa podmienkam a prežiť, o to viac by v rovine intelektuálnej prebiehal zápas medzi ľudskou civilizáciou a limitovanou AGI. **Prekročenie koncepčných prelomov a následná inteligenčná explózia, o ktorej budeme uvažovať v kapitole 5.4.11., by pre AGI bola príliš veľkou „evolučnou“ výhodou v porovnaní s človekom**.

Na akýkoľvek biologický ekosystém môžeme nazerať ako na „kvázi“ organizmus, ktorý obsahuje emergentnú vlastnosť, t.j. určité mechanizmy stability, ktoré zabezpečujú, že tento systém je udržateľný a dlhodobo funguje. Mechanizmy stability emergentného systému sú rezistencia (odolnosť) a resiliencia (adaptabilná pružnosť).

Analogicky v digitálnej ére budujeme komplexnú dátovú, informačnú a technologickú infraštruktúru, ktorá sa v budúcnosti môže pre pokročilé systémy AI stať emergentným ekosystémom, umožňujúcim bezproblémovú funkčnosť kooperujúcich adaptívnych a autonómnych systémov umelej inteligencie.[603] **Miera stability emergentného ekosystému AI by mohla byť problémom pre našu schopnosť ho ovládať a držať pod kontrolou v scenároch, v ktorých by nám AGI začala prerastať cez hlavu**.

Ďalším rizikom by bolo zvýšená odolnosť emergentného ekosystému umelej inteligencie pri stanovovaní a dosahovaní inštrumentálnych cieľov. Ako sme uvádzali v kapitole 5.4.9., prostredníctvom inštrumentálnych cieľov dokážu systémy limitovanej AGI zabezpečiť potrebné mechanizmy stability svojej činnosti pri dosahovaní konečných cieľov. Rezistencia a resiliencia biologických emergentných ekosystémov sa nápadne podobá inštrumentálnym cieľom ochrany pred vypnutím (strachu) a zabezpečovania dostatočných zdrojov (chamtivosť).

Evolučný vývoj a zdokonaľovanie, resp. vznik nových druhov je odpoveďou na zmenu

602 Ako sme uvádzali v kapitole 5.2., vo vývoji technológií AI existuje snaha niektorých kruhov aplikovať evolučný vývoj na systémy AGI – napr. evolučné genetické polymorfné algoritmy a reinforcement learning využité za špecifických podmienok (učenie sa vo veľmi komplexnom prostredí – dataset na úrovni reálneho sveta) na smerovanie k Life 3.0 (ASI).

603 Ide o jeden z predpokladaných aspektov AGI, keď by inteligentné stroje neboli osamotenými systémami, ale mali by globálny dosah.

podmienok v prírode. Ďalším fenoménom v evolúcii je konkurencia, ak je viacero druhov, ktoré „robia to isté“, majú šancu sa prispôsobovať a ďalej vyvíjať, pričom len tie najúspešnejšie druhy preberajú pomyselné žezlo vývoja a budúcej existencie.

Analogicky sa môžeme pozerať na – obrazne povedané – inteligentné druhy ľudská bytosť vs. AGI. Ak by sme nezvládli vývoj technológií limitovanej AGI a kontrolu nad nimi, mohli by sme sa stať súčasťou konkurenčného boja, v ktorom by však väčšinu výhod na svojej strane mali inteligentné stroje.

Analogicky k evolučnému vývoju, ktorý je stimulovaný zmenami v prostredí – pokiaľ nemáme adekvátny model správania a teóriu mysle AI, nevieme ako digitálne stimuly virtuálneho sveta dokážu v konečnom dôsledku ovplyvniť a zmeniť činnosť AGI. Zvládnutie koncepčných prelomov AI prináša veľa stupňov voľnosti[604] vo vývoji a činnosti inteligentných strojov pri dosahovaní konečných cieľov – t.j. existuje riziko, že možné reakcie systémov limitovanej AGI na prostredie v kontexte dosahovaných konečných cieľov môžu byť nepredvídateľné.

5.4.11. Explózia inteligencie strojov ako problém pre ľudstvo

Pod explóziou inteligencie myslíme proces seba vylepšovania všeobecnej umelej inteligencie, ktorý by mohol viesť až k superinteligencii (ASI)[605], t.j. inteligencii, ktorá by bola naprieč všetkými oblasťami inteligentnejšia ako človek. Úvahy o ASI sa mnohokrát spájajú s konceptom singularity v oblasti AI, v ktorej umelá inteligencia so schopnosťou učiť sa, zlepšovať sa a samostatne sa vyvíjať rýchlo dosiahne a následne extrémne prevýši inteligenciu človeka, pričom by mohlo ísť o exponenciálny proces.[606]

Ak by systém AGI svojou inteligenciou ďaleko presahoval inteligenciu najmúdrejších ľudí a súčasťou jeho inteligencie by bezpochyby bola aj intelektuálna aktivita zameraná na vylepšovanie dizajnu inteligentných strojov, mohol by vytvoriť ešte lepšiu AGI. To by viedlo k špirále inteligenčnej explózie, ktorej výsledkom by

---

604 Ide o alúziu na Bostromovu *tézu o nezávislosti* (the orthogonality thesis), ktorú uvádzame v kapitole 5.5.1.: „Inteligencia a konečné ciele sú nezávislé: viac-menej akúkoľvek úroveň inteligencie možno v zásade kombinovať s viac-menej akýmkoľvek cieľom.“

605 ASI – Artificial Super Intelligence.

606 V kontexte nášho rozlišovania tzv. slabej a silnej singularity AI, v prípade ASI hovoríme o možnosti silnej singularity AI.

bolo zanechanie človeka ďaleko za sebou.[607]

V kontexte doteraz uvedených rizík a problémov je jasné, že inteligenčná explózia môže viesť k strate ovládania a kontroly nad takýmto systémom umelej inteligencie.

Prvý systém AGI by tak mohol byť posledným ľudským výtvorom, ktorý by človek musel ešte spraviť. Samozrejme za predpokladu, že tento systém by bol natoľko poddajný, že by nám prezradil, ako ho udržať pod kontrolou. Znovu, i v tomto kontexte musíme podotknúť, že ak by – úplne hypoteticky – limitovaná AGI schopnosti superinteligencie dosiahla, museli by sme o nej uvažovať ako o niečom, čo je inteligentné úplne iným spôsobom než ľudská bytosť.

Na základe inštrumentálnych cieľov, medzi nimi tých, ktoré sme obrazne nazvali strach (snaha chrániť sa pred vypnutím) a chamtivosť (snaha o zabezpečenie potrebných prostriedkov pre svoju činnosť), môžeme predpokladať, že systémy AGI by nielenže mohli, ale by aj skutočne podnikli kroky k seba vylepšovaniu smerujúc tak k superinteligencii. Bolo by im totiž úplne jasné, že tieto vylepšenia by priniesli mnohé benefity na ceste k dosahovaniu stanovených cieľov a zadaných úloh.[608]

Exponenciálna trajektória týchto seba vylepšujúcich krokov v zásade znamená, že by išlo o veľmi rýchly proces, ktorý by ľudstvu nedával veľa času na vyriešenie problému ovládania takejto inteligencie.[609]

V hypotetickom náhľade na možné dôsledky inteligenčnej explózie sa nám naskytá ešte jedna alternatíva – i keď sa na to nemôžeme spoliehať, musíme rátať aj s eventualitou klesajúcej návratnosti vylepšovania inteligencie, ktorá môže mať až taký regres, že sa vylepšovanie v určitom bode miesto explózie zastaví a vyčerpá.[610] Inak povedané, percentuálne vylepšenie sa by sa v určitom bode špirály narastajúcej

607 GOOD, I. J. *Speculations concerning the first untraintelligent machine*. In: *Advances in Computers*. Academic Press, vol. 6, 1965.

608 Por. RUSSELL, *Human Compatible*, s. 143.

609 Nick Bostrom tento proces prezentuje ako „scenár tvrdého rozletu“, pri ktorom inteligencia strojov astronomicky vzrastie v priebehu dní alebo týždňov.
Tento i ďalšie rizikové faktory a dopady inteligenčnej explózie rozoberá Luke Muehlhauser z MIRI (Machine Intelligence Research Institute), pričom nadväzuje na obavy takých osobností, ako sú Alan Turing, I. J. Good,... Bill Joy and Stephen Hawking.
MUEHLHAUSER, L. *Facing the Intelligence Explosion*. MIRI, 2013.

610 YUDKOWSKY, E. *Intelligence explosion microeconomics*. MIRI, 2013.

inteligencie mohlo stať oveľa náročnejšie, než za túto cenu dosiahnutý pokrok a posun v strojovej inteligencii.

V súčasnosti nie sme v stave, že by sme dokázali túto eventualitu vylúčiť.

Logickým dôsledkom uvedeného rizika zastavenia inteligenčnej explózie na základe vyčerpania procesu sebazdokonaľovania strojovej inteligencie je poznanie, že **ak AGI nebude schopná seba vylepšovania až po úroveň superinteligencie, nebudú toho schopní ani ľudia**.

Na základe nami predloženého konceptu limitovanej AGI by obmedzujúcim faktorom, ktorý systémy AI v seba vylepšovaní nedokážu prekročiť, mohla byť aj hranica medzi inteligenciou a vedomím, resp. sebauvedomením, ktoré v spojení s rozumom a slobodnou vôľou chápeme ako mohutnosti duše, presahujúce výlučne biologickú realitu mozgu a nervovej sústavy. Išlo by tak o problém, ktorý z podstaty veci by sme ani my a ani seba rozvíjajúca umelá inteligencia nedokázali vyriešiť a túto hranicu prekročiť.

Než by vylepšená AGI začala byť nebezpečenstvom pre ľudstvo, v kontexte predikovaných možností a potencionálnych ohrození ľudskej civilizácie by však bolo veľmi riskantné spoliehať sa, že scenár vyčerpania seba vylepšovania pokročilej strojovej inteligencie nastane, alebo že nastane dostatočne včas, prípadne že jeho bariérou a definitívnou stopkou bude absencia vedomia a sebauvedomenia u limitovanej AGI.

Ak by však systémy AGI predsa len boli schopné seba vylepšovania a inteligenčnej explózie, a ak by sme neboli schopní vyriešiť problém ovládania čo i len mierne „nadľudskej“ AGI – napr. ak by sme neboli schopní zastaviť, resp. ovládať špirálu rekurzívneho seba vylepšovania – ako ľudstvo by sme už nemali žiadny čas a priestor na vyriešenie problému riadenia a ovládania AGI a v zásade by sme skončili.[611]

5.4.12. Môže nám limitovaná AGI skutočne prerásť cez hlavu?

Môžeme namietať, že limitovaná AGI zo svojej podstaty nikdy nedosiahne úrovne inteligencie človeka v kontexte vedomia a sebauvedomenia, no nesmieme zabúdať na v tejto kapitole viackrát uvedené dôsledky rizík sofistikovanej umelej inteligencie na úrovni limitovanej AGI, ktorými môže byť postupná hodnotová a psychologická degradácia ľudského bytia spojená so sociálnou a ekonomickou disrupciou a premenou

611 Povedané slovami Alana Turinga: “je to určite niečo, čo by nám navodilo úzkosť”.

spoločnosti. Inými slovami – **existuje riziko, že miesto osobného a civilizačného rastu vďaka benefitom prameniacim z využívania a ovládania limitovanej AGI, nám môže hroziť riziko úpadku**.

S civilizačným pokrokom sa človek i celé ľudstvo dvíha zo zeme a stojí na vrchole pyramídy stvoreného sveta. To, čo v oblasti umelej inteligencie chceme vytvoriť, môže mať potenciál postaviť sa na niektorý zo stupňov tejto pyramídy.

Pokiaľ v rámci rozvoja a využívania AGI dokážeme zabezpečiť, že by si ľudstvo zachovalo dostatočnú autoritu nad strojmi, chápanie ich procesov a svoju vlastnú autonómnosť, tieto systémy môžu pomôcť znásobiť ľudské schopnosti a umožniť nám postaviť sa na ešte vyššie stupne pyramídy rozvoja.[612]

Ak to nedokážeme, naše miesto v pyramíde zostane zachované, no nemusí byť už jej vrcholom, keďže budeme len súčasťou pyramídy, v ktorej budú nad nami iné stupne – stupne strojovej inteligencie, pričom bude len otázkou času, kým nás zavanie prach…[613]

5.5. Čo robiť, aby sme nezapadli prachom

Javí sa, že na riziká a úzkosť, ktorú s možným nástupom superinteligencie pociťoval Turing a po ňom až po súčasnosť mnohí ďalší, je možné reagovať viacerými spôsobmi, medzi ktoré patrí:[614]

- **ústup od výskumu všeobecnej a silnej umelej inteligencie**
- **rezignácia – prenechanie budúcnosti inteligentným strojom**
- **popieranie rizík spojených s vývojom pokročilej umelej inteligencie**
- **pochopenie a zmiernenie rizík prostredníctvom návrhu systémov umelej**

612 Por. RUSSELL, *Human Compatible*, s. 131.

613 Spomeňme si na Hofstadterove obavy vyjadrené v predslove tejto knihy, ktorými v centrále Google reagoval na Kurzweilovu predikciu ohľadom strojovej superinteligencie: „Bol som vydesený týmto scenárom. Veľmi skeptický, no v rovnakom čase, mysliac si, že i keď predpokladaný časový harmonogram je nereálny, možno majú pravdu. A potom budeme úplne zaskočení. Budeme si myslieť, že sa nič nedeje a zrazu, skôr než si to uvedomíme, budú počítače múdrejšie než my.“ A keď sa to stane, „budeme nahradení, staneme sa reliktom. Zostaneme v prachu.“ Končiac svoj príhovor dodáva: „Možno sa to stane, ale nechcem, aby sa to stalo priskoro. Nechcem, aby moje deti zostali v prachu.“ Por. MITCHELL, *Artificial Intelligence*, s. 10-11.

614 RUSSELL, *Human Compatible*, s. 144.

inteligencie, ktoré nevyhnutne zostanú pod ľudskou kontrolou

Konkrétne dôvody pre **nereálnosť ústupu** od výskumu všeobecnej umelej inteligencie sme uvádzali v kapitole 5.4.7, venujúc sa problému goríl.

Na rozdiel od ústupu od výskumu je rezignácia možnou, ale asi tou najhoršou možnou reakciou. V namiešanom kokteile postmodernistických naratívov alternatívu rezignácie často sprevádza myšlienka, že **systémy umelej inteligencie, ktoré by boli inteligentnejšie ako my, si nejakým spôsobom zaslúžia zdediť planétu a nechať ľudí v pokoji odísť do prítmia minulosti utešujúc sa myšlienkou, že naši geniálni elektronickí potomkovia sú zaneprázdnení dosahovaním svojich super cieľov**. Tento názor hlásal robotik a futurista Hans Moravec: „Nesmrteľnosť kybernetického priestoru sa bude hemžiť neľudskými 'supermúdrosťami' zaoberajúcimi sa záležitosťami, ktoré sú pre ľudské záujmy také, ako sú tie naše pre záujmy baktérií.“[615]

Z pohľadu zložitosti logických procesov prebiehajúcich v hypotetických kybernetických super mysliach by Moravec mohol mať aj pravdu. Avšak **v kontexte hodnôt a schopností, ktoré v našom ponímaní viažeme na vedomie a sebauvedomenie, by stále mohlo ísť o technologický „balast“, ktorý bez hodnotového rámca a vyššieho zmyslu nie je ničím. A je potom na zamyslenie, či práve vyšší zmysel z nás nerobí to, kým sme a neumožňuje nám kráčať cestou civilizačného rozvoja túžiaceho sa dotknúť transcendentna**.

5.5.1. Odborná debata, ktorej niečo chýba

Podľa toho, čoho sme za uplynulé desaťročia vývoja umelej inteligencie svedkami, popieranie rizík spojených s vývojom pokročilej AI môže byť realizované rôznymi spôsobmi a v rámci nich môže nadobúdať i rozličné formy. Naviac – okrem klasického popierania rizík spojených s vývojom AGI – sa stretávame i s cieleným odklonom od tejto témy a zľahčovaním celej problematiky.

V rámci popierania problému rizík spojených s vývojom AGI sa inteligencia

615 MORAVEC, *Mind Children: The Future of Robot and Human Intelligence*.
MORACEC, H. *Robot: Mere Machine to Transcendent Mind*. Oxford University Press, 2000.
Prof. Russell túto úvahu považuje za chybnú, keďže „hodnota je pre ľudí definovaná predovšetkým vedomou ľudskou skúsenosťou. Ak neexistujú ľudia a iné vedomé entity, na ktorých subjektívnej skúsenosti nám záleží, nič hodnotné sa nedeje.“
RUSSELL, *Human Compatible*, s. 144.

analogická ľudskej považuje za natoľko komplikovanú, že v plnosti sa jej stroje nedokážu vyrovnať. Tento argument neobstojí, keďže neriešime riziká vyplývajúce z inteligencie, ktorá by bola kópiou tej ľudskej, ale z výziev, ktorých súčasťou sú rizikové faktory viažuce sa na algoritmy inteligencie umelej a ako sme už skôr viackrát deklarovali, pravdepodobne diametrálne odlišnej.

Ďalšou formou popierania je vyvracanie možnosti vytvorenia AGI, resp. vyvracanie existencie rizík, ktoré sme rozoberali v kapitole 5.4. Na základe toho, že doteraz a v rámci súčasného boomu AI práve teraz sa nám nedarí vytvoriť AGI, tak riziká s ňou spojené vraj nie sú reálne.[616] Je určitým paradoxom, že popieranie možnosti vytvorenia limitovanej AGI je prezentované inštitúciami, ktoré sú súčasťou špičkového vývoja v tejto oblasti a pravidelne prezentujú ďalšie a ďalšie úspechy pri prekračovaní bariér na ceste k všeobecnej umelej inteligencii.[617]

Vytvorenie AGI ešte dlho nebude možné – tak znie pozmenená, resp. komplementárna výhrada k nemožnosti aktuálne dosiahnuť technológie všeobecnej umelej inteligencie. V zásade ide o argumentačný faul – vytvárajúc si tzv. *strašiaka* (straw man) aktuálneho ohrozenia, s ktorým by sme sa mali vysporiadať. Avšak už od kapitoly 5.3. riešime nie aktuálne ohrozenie nejakou z ničoho nič sa vynoriacou superinteligenciou, ale serióznu možnosť dosiahnutia koncepčných prelomov pri vývoji AI a z toho prameniace riziká, ktorým dokážeme čeliť len vtedy, ak už teraz máme premyslenú stratégiu vývoja a dizajnu týchto prelomových systémov.

Korunou popierania je však elitárstvo a kmeňová príslušnosť. Odborníci na umelú inteligenciu samozrejme oveľa hlbšie chápu problematiku vývoja algoritmov AI a z toho plynúce konsekvencie pre človeka a spoločnosť. Avšak práve na základe svojej odbornosti by mali byť schopní viesť debatu na relevantnej odbornej úrovni a na základe serióznych argumentov, a nie skĺznutím do roviny argumentácie *ad hominem*. Miesto zamyslenia sa, argumentácie a vysvetľovania sú výhrady oponentov apriori prezentované ako

616 Je to analogické príkladu s vyriešením problému štiepenia jadra, ktorý sme uvádzali v kapitole 5.3 – tvrdenie Ernesta Rutherforda o nemožnosti vyriešenia tejto úlohy, ktoré však vďaka Leóvi Szilárdovi behom dvadsaťštyri hodín ľahlo popolom.

617 *One Hundred Year Study on Artificial Intelligence*. [on-line]. [cit. 15. novembra 2021]. Dostupné na internete: <https://ai100.stanford.edu/2021-report>
Paradoxom v tomto popieraní je, že prestížna stanfordská štúdia, ktorú sme už skôr v tejto publikácii spomínali, zastrešuje aj inštitúcie a vedcov, ktorí sa o dosiahnutie AGI aktívne usilujú.

spiatočnícke a tmárske,[618] pričom odbornosť ich nositeľov je znevažovaná.[619] Určite existuje nemálo prípadov, kedy tomu tak je, no skutočná podstata problému zostáva a prezentujú ju nie nepriatelia rozvoja modernej civilizácie, ale protagonisti seriózneho, bezpečného a na dobro človeka orientovaného pokroku a vývoja AI. Protagonisti, z ktorých viacerých sme už spomínali, pričom väčšina z nich patrí k špičke či už v oblasti umelej inteligencie a robotiky (Turing, Wiener, Hofstadter, Bostrom, Good, Russell,...), alebo v oblastiach, ktoré svojím dielom svoj odbor prerástli a stali sa osobnosťami, nad ktorých názorom sa treba zamýšľať (Hawking, Gates, Musk,...).

Tiež sa stáva, že akákoľvek snaha o reálnu diskusiu a riešenie nebezpečenstiev je proklamovaná ako útok na výskum a snahu ho zakázať, či neadekvátne obmedziť. Ide o falošnú predstavu, že ktokoľvek, kto hovorí o rizikách AI, ignoruje jej potenciálne výhody, ba dokonca ich neguje.[620] Táto úvaha je chybná z dvoch dôvodov:[621]

- ak by neexistovali potencionálne výhody umelej inteligencie, neexistoval by ekonomický a spoločenský dopyt po výskume v oblasti AI, takže by ani neexistovalo nebezpečenstvo z dosiahnutia AGI;
- pokiaľ by riziká AI neboli úspešne zmiernené, resp. eliminované, nebolo by možné participovať na jej benefitoch.

V prípade vývoja všeobecnej umelej inteligencie ide o tak vážne dôsledky na život a existenciu človeka i fungovanie a budúcnosť spoločnosti, že argument „sme

618 Napr. Oren Etzioni, riaditeľ Allen Institute for AI a známy vedec v oblasti ML a NLP, kladie argumenty a obavy seriózných vedcov na roveň bezhlavého nepochopenia a strachu z nových technológií, pričom častokrát ide o argumenty tých, ktorí sa o vznik daného odboru zaslúžili, alebo k jeho súčasnému rozvoju a stavu poznania v nemalej miere prispeli:
„Pri vzniku každej technologickej inovácie sa ľudia báli. Od tkáčov, ktorí na začiatku priemyselnej éry hádzali topánky do mechanických krosien, až po dnešný strach zo smrtiacich robotov, bola naša reakcia spôsobená nevedomosťou, aký vplyv bude mať nová technológia na náš život a jeho zmysel. A keď to nevieme, naša ustráchaná myseľ si doplní detaily.“
ETZIONI, O. *It's time to intelligently discuss artificial intelligence*. Backchannel, Dec 9, 2014.

619 Napr. SOFGE, E. *Bill Gates fears AI, but AI researchers know better*. Popular Science, Jan 30, 2015.

620 Na základe takejto argumentácie by museli odstaviť 80-90% vedcov, ktorí sa podieľali na vývoji atómovej bomby v Los Alamos a ktorí rozoberali reálne konsekvencie svojho výskumu. A v súčasnosti by sme museli zrušiť všetky regulácie výskumu a vývoja v oblasti jadrových zbraní a atómovej energie, biotechnológií, ochrany osobných údajov, priemyselných noriem a pod.

621 RUSSELL, *Human Compatible*, s. 157.

experti, vieme to lepšie než vy a máme to pod kontrolou“ je absolútne neprijateľný a výskum by mal byť predmetom serióznej odbornej diskusie a prípadnej regulácie po vzore iných odvetví s analogickým dopadom na ľudskú civilizáciu.

Uvedomujeme si, že v súčasnej dobe, keď sa odborná diskusia v mnohých oblastiach nahrádza plytkým aktivizmom a zaniká pod ideologickým, politickým či lobistickým tlakom, nie je jednoduché túto odbornosť i serióznu úroveň diskusie zachovať a ísť s „kožou na trh“. **Súčasne sa aktéri diskusie na základe elitárstva alebo názorových bublín môžu postupne viac a viac stávať súčasťou kmeňovej príslušnosti, ktorej ovocím je miesto racionálnej debaty nezmieriteľná a stagnujúca zákopová vojna**.[622] Vzhľadom na závažnosť rizík spojených s AGI však neslobodno túto odbornú debatu a hľadanie riešení zadusiť v strachu z nepredvídateľných a neodborných zásahov nátlakových skupín, alebo v oslepení bojovými kmeňovými farbami, do ktorých sme vedome či nevedome ponorili naše mysle i slová.

Cielený odklon od témy je často prepletený so zľahčovaním celej problematiky, pričom debata sa viac či menej vzďaľuje od riešenia rizikových faktorov AGI.

Napríklad problém s nemožnosťou vypnúť systém AGI vzhľadom na inštrumentálny cieľ „strachu“ (kapitola 5.4.9) sa odbaví uistením, že robustnosť systému by bola považovaná za chybu a ako taká by bola vo vývoji eliminovaná. Môžme síce debatovať o tom, či predpokladaná neuveriteľná robustnosť by bola požadovanou vlastnosťou alebo chybou, každopádne superinteligentná AI ju tak či tak môže (a takmer naisto tak i spraví) použiť na svoju vlastnú ochranu. V tomto prípade ide o odklon od riešenia problematickej činnosti inteligentného systému, ktorý nie sme schopní vypnúť a zľahčovanie jeho schopností, v rámci ktorých by takýto systém na základe dosiahnutých koncepčných prelomov nebol odkázaný na implementáciu, resp. blokovanie robustnosti, keďže sám by si túto schopnosť dokázal doplniť.

Jednou z povšimnutia hodných pokusov o riešenie je snaha AGI izolovať a prístup

622 Symptómy tejto kmeňovej príslušnosti poznáme: vzájomná nedôvera a očierňovanie, iracionálne argumenty a odmietanie uznať akýkoľvek (rozumný) názor, ktorý by mohol byť v prospech iného kmeňa. Na strane zástancov technológie vidíme popieranie a zatajovanie rizík v kombinácii s obvineniami zo spiatočníckeho tmárstva. Na strane odporcov vidno presvedčenie, že riziká sú neprekonateľné a problémy neriešiteľné. Zmierlivé hlasy na obidvoch stranách sú potláčané a považované za zradné, takže slovo majú len dôrazne prezentované extrémne názory, ktoré sú druhou stranou neprijateľné. RUSSELL, *Human Compatible*, s. 159-160.

k nej umožniť len ako filtrované vstupy a výstupy, ktoré by sme mali pod kontrolou. Systém, ktorý by v tomto režime fungoval, dostal poetický názov Oracle AI – všeobecná umelá inteligencia, s ktorou by sme narábali ako s takmer vševediacou veštieckou skrinkou.

Oracle AI však v sebe skrýva dva problémy:[623]

- kvalita vstupných dát, ako jeden z extrémne rizikových faktorov AI, ktorý sme uvádzali v kapitole 2.1.2. Oracle AI bude len natoľko správne fungujúca, nakoľko jej pripravíme dostatočné a správne vstupné údaje, a tie budú limitované ľudským faktorom na vstupe (resp. predprípravou oveľa menej inteligentného stroja).
- dostatočne inteligentná Oracle AI sa naučí svojimi odpoveďami ovládať obslužný personál. Na základe skutočností, ktoré sme uviedli v kapitole 5.3.1, nie je ťažké uhádnuť, ako by asi takéto „väzenie“ skončilo – na základe dôsledkov vyriešenia koncepčných prelomov by sa z neho AGI s našou pomocou skôr či neskôr dostala.[624]

Na druhej strane treba uznať, že Oracle AGI by bola ovládateľnejšia než AGI, resp. ASI. Preto by možno bolo vhodné na priamej ceste k AGI skôr realizovať Oracle AGI a postupne riešiť ešte nevyriešené rizikové faktory AGI, o ktorých už vieme, alebo riziká a výzvy, ktoré z dnešného uhľa pohľadu nedokážeme odhaliť.[625]

Veľa nádejí sa vkladá do možnosti pracovať v spoločných tímoch tvorených ľuďmi a inteligentnými strojmi. Tímová spolupráca ľudí s AGI je určite žiadaným cieľom, problémom sú však ciele, ktoré sa môžu líšiť. Preto pomenovanie problému rôznorodosti a potreby zjednotenia cieľov je síce fajn, ale neznamená to, že na tento problém máme riešenie.

Ide tak o jeden z klasických odklonov od témy rizík, akcentujúc očakávané – a treba

---

623 Por. RUSSELL, *Human Compatible*, s. 162.

624 Niektoré dialógy z filmu Ex Machina medzi humanoidnou umelou inteligenciou Ava, jej tvorcom Nathanom a testerom Calebom sú možno jednoduchou, ale názornou ukážkou, ako by humanoidná AI dokázala zneužiť riziká, o ktorých sme uvažovali v kapitole 5.4.6 a zmanipulovať ľudský personál pre dosiahnutie svojich zámerov.
Ex Machina (film). [on-line]. [cit. 30. septembra 2022].
Dostupné na internete: <https://en.wikipedia.org/wiki/Ex_Machina_(film)>

625 RUSSELL, *Human Compatible*, s. 162.

uviesť, že i žiadané – benefity z nasadenia systémov AGI so súčasným ignorovaním, prípadne zľahčovaním existujúcich rizík. Tento prístup sa niekedy prejavuje až vo svojej extrémnej forme – o rizikách spojených s AGI jednoducho mlčať a polemiku okolo nich ignorovať.

Ďalším z navrhovaných riešení je prepojenie ľudského mozgu a AGI do jednej rozšírenej, resp. vylepšenej mysliacej entity s vlastným vedomím, ktorá by bola pod kontrolou ľudského mozgu, t.j. človeka. Tento koncept je jednou z možných aplikácií limitovanej AGI zasahujúcej do transhumanizmu, ktorá by riešila i vylepšovanie človeka i jeho ochranu pred pokročilou umelou inteligenciou.[626]

Odhliadnuc od bioetických výhrad, nutnou podmienkou tohoto riešenia je vytvorenie robustného a permanentného prepojenia (centier) ľudského mozgu a externého výpočtového systému, resp. kybernetickej siete. V súčasnosti však existujú dva problémy, bez vyriešenia ktorých to nepôjde.[627]

Prvým, pravdepodobne riešiteľným, je neexistujúca relevantná technológia napojenia elektronického systému na mozgové tkanivo.[628]

Druhým, ktorý je aj v pohľade do budúcnosti veľkou neznámou, je naša súčasná minimálna znalosť a takmer nulové chápanie neuralgickej implementácie vyšších kognitívnych schopností mozgu, bez čoho je takmer nemožné správne sa na mozgové centrá napojiť a realizovať relevantnú procesnú interakciu medzi mozgom a strojom.

Určite v tomto riešení nenachádzame elimináciu rizík prameniacich zo schopností limitovanej AGI, nakoľko už na základe využívania prostriedkov slabej umelej inteligencie evidujeme posuny v ľudskom intelekte zahŕňajúce pozmenené vnímanie a kognitívne procesy, emocionálne rozpoloženie, labilitu a nechcené psychické stavy i nezanedbateľnú mieru manipulácie a ovplyvňovania, ako sme uvádzali v kapitolách 2.5 a 5.4. Neexistujú garancie, na základe ktorých by sme mohli s istotou povedať, že ľudský rozum a vôla by

626 V zásade ide o dva základné prieniky medzi umelou inteligenciou a transhumanizmom, ktoré sme uvádzali v kapitole 1.10.

627 RUSSELL, *Human Compatible*, s. 164.

628 Už v úvode tejto publikácie sme spomínali Neuralink – rozhranie medzi mozgom a počítačom pre rozšírenie možností človeka, ako napr. integrácia robotických končatín, alebo prepojenie ľudského mozgu a umelej inteligencie pre dosiahnutie symbiózy. Ide o projekt Elona Muska, ktorý rozvíja v rámci svojich vizionárskych aktivít (Neuraling Corporation).

boli rozhodujúcim a podstatným faktorom takejto symbiózy. Nezanedbateľný aspektom je i transhumanistický posun, ktorý sa dotýka podstaty človeka, jeho identity a hodnotového rámca.

Každopádne, ako poznamenáva prof. Russell, ak by sme toto spojenie ľudského mozgu a strojovej inteligencie skutočne potrebovali, aby sme v konfrontácii s naším výtvorom – AGI prežili, asi sme niekde urobili chybu.[629]

Prakticky celá doterajšia rozprava o rizikách všeobecnej umelej inteligencie sa dotýkala problematiky požadovaných cieľov činnosti AI a pre ich dosiahnutie aj adaptívne realizovaných parciálnych, resp. inštrumentálnych cieľov, ktoré – ako sme videli – by mohli spôsobovať ľudstvu veľké problémy, keďže ich nedefinujeme a neusmerňujeme my, ale si ich autonómne stanovuje samotná AGI. Umelá inteligencia si ich nielen sama stanovuje (napr. požiadavku na zdroje alebo ochranu pred vypnutím), ale ich obsah a spôsob pochopenia si aj sama definuje.

Napríklad, ak by pre svoju činnosť systém AI potreboval narábať s konceptom usmrtenia živého tvora, ktorý by stál v ceste uskutočnenia stanoveného cieľa, pre AI by smrť bola len spôsobom, ako zastaviť fungovanie „subjektu“ a odstrániť prekážku dosiahnutia cieľa. Absolútne by nebola schopná vnímať morálny rozmer usmrtenia. Ak by tento koncept rozšírila na usmrtenie vo všeobecnosti, pre stroj by to nebolo morálne zlo (ukončenie života,...), ale len technická strata alebo výhoda na ceste k dosiahnutiu cieľa. Možno naviac s chladnou kalkuláciou dopadu tejto udalosti na postoje ľudí, ktorí boli v nejakom vzťahu k usmrtenej bytosti, prípadne prekážky, ktoré by sa mohli v kontexte zákonov spoločnosti objaviť...

Vo všeobecnosti sa môže zdať, že rizikové faktory fungovania AGI sú dôsledkom špecifických druhov cieľov, ktoré ak by sme odstránili, riziká by pominuli.[630]

Preto ako najjednoduchšie riešenie sa javí zablokovanie rizikových cieľov, resp. obmedzenie cieľov zadávaných strojovej inteligencii (a tým aj eliminovanie vytvárania jej inštrumentálnych cieľov, t.j. parciálnych krokov potrebných na dosiahnutie zadaných cieľov, ktoré by taktiež mohli byť rizikové). Avšak akýmkoľvek obmedzením cieľov, či už parciálnych (ak by sme to vôbec dokázali) alebo celkových, systém AI degradujeme na deterministický stroj bez vlastnej autonómie a adaptačných schopností. Miesto strojovej

629 RUSSELL, *Human Compatible*, s. 165.

630 RUSSELL, *Human Compatible*, s. 165.

inteligencie tak budeme mať len ďalší „nadupaný“ počítač.

Druhá možnosť – systémom AGI nezadávať ľudské ciele, veď pokročilá umelá inteligencia si ich dokáže stanoviť sama – je ešte horšia a ide v ústrety hororovým scenárom sci-fi. **Bez predstavy, že na ľudských preferenciách záleží a bez cieľov, ktoré určujú ľudia, neexistuje dôvod, aby pokročilá strojová inteligencia konala v prospech ľudí**.[631]

Na základe niektorých fáz vývoja ľudskej spoločnosti[632] mnohí prezumujú, že u ľudí s vyššou inteligenciou sa prejavuje tendencia mať altruistické a vznešené ciele, ktoré naviac môžu byť nachádzané a samostatne stanovované v prostredí reálneho sveta.

V určitej forme sa touto hypotézou už v polovici 18. storočia zaoberal filozof David Hume (nazýval ju „is-ought“ problem), pričom prišiel k záveru, že je chybou myslieť si, že morálne imperatívy možno odvodiť z prírodných faktov.[633]

K podobnému záveru prichádza i Nick Bostrom v už spomínanom diele Superinteligencia a dáva mu svoje vlastné pomenovanie *téza o nezávislosti* (the orthogonality thesis): „**Inteligencia a konečné ciele sú nezávislé: viac-menej akúkoľvek úroveň inteligencie možno v zásade kombinovať s viac-menej akýmkoľvek cieľom**.“[634]

Podľa prof. Russella ide o tak zásadnú tézu, že pre inžinierov a počítačových vedcov zaoberajúcich sa štandardným modelom inteligentných strojov a ich cieľov je jednoducho daná a patrí k základným prerekvizitám ich práce. Nie je možné predpokladať, že by inteligentné systémy mohli len z pozorovania sveta vyextrahovať ciele svojej činnosti. Naviac také ciele, ktoré sú správne, takže dostatočne inteligentný algoritmus by sa prirodzene vzdal svojho pôvodného cieľa v prospech „správneho“ cieľa.[635]

Na druhej strane by limitovaná AGI mohla chápať, že pre ňu optimálny plán na dosiahnutie

---

631 Povedané slovami Nicka Bostroma: „bez cieľov totiž neexistuje dôvod, aby stroj uprednostnil ľudský raj pred planétou premenenou na more kancelárskych spiniek.“ Tento motív sa nesie takmer celým jeho dielom Superinteligencia.
BOSTROM, *Superintelligence*, s. 107-243.

632 Por. FLYNN, *What Is Intelligence?: Beyond the Flynn Effect*.

633 V svojom Pojednaní o ľudskej povahe sa Hume venoval viacerým filozofickým otázkam, medzi ktoré patril i problém, či možno morálne záväzky vypozorovať z prirodzeného sveta.
HUME, D. *A Treatise of Human Nature*. John Noon, 1738.

634 BOSTROM, *Superintelligence*, s. 107.

635 Por. RUSSELL, *Human Compatible*, s. 167-168.

konečného cieľa by mohol ľuďom spôsobiť problémy, no – ako sme skôr poznamenali – už z princípu to v optike splnenia svojho cieľa nebude považovať za nedostatok a nebude sa tým ani zaoberať, pokiaľ to nebude mať priamo zadané.

Ak toto všetko porovnáme s modelom správania človeka, na rozdiel od AI sa my, ľudia apriori (z rôznych dôvodov) staráme aj o preferencie iných ľudí a dokonca kalkulujeme s tým, že ani nemusíme poznať všetky tieto preferencie a dôsledky nášho konania na iných. Z toho pramení aj spôsob, ako konáme, aby – i bez presne zadaných cieľov – sme tieto preferencie pokiaľ možno neskrížili a konali korektne voči druhým.[636]

Vnímajúc, že popieranie a zľahčovanie rizík spojených s vývojom technológií pokročilej umelej inteligencie netvorí seriózny prístup k problémom, ktoré by mohla limitovaná AGI pre jednotlivca i celú modernú spoločnosť znamenať, **nezostáva nám nič iné, len sa snažiť rizikové faktory pochopiť a hľadať cesty, ako zmierniť obavy z ohrozenia, ktoré by všeobecná a silná AI mohla pre ľudskú civilizáciu predstavovať. Obnáša to rast povedomia o rizikách skrývajúcich sa za koncepčnými prelomami AI, seriózne vedecké zameranie sa na riešenie týchto rizík, politickú podporu pre potrebné riešenia (napr. regulácie, osveta,...) i alokovanie dostatočných zdrojov**.

Čím viac a serióznejšie sa bude vývoj umelej inteligencie venovať rizikám prameniacim z koncepčných prelomov, tým reálnejšia je šanca, že dokážeme dospieť k takému návrhu systémov AGI, ktorý si zachová „dobré zvyky ANI“, t.j. stále pôjde o umelú inteligenciu, ktorá dokáže byť orientovaná na dobro človeka.

5.5.2. Základné východiská pre riešenie

Uvedomujeme si, že na základe dosiahnutia a prekročenia koncepčných prelomov je limitovaná AGI (t.j. podľa nás jediná možná realizovateľná podoba všeobecnej umelej inteligencie[637]) z pohľadu bezpečnostných atribútov, etických výziev, morálnych aspektov a celkového dopadu svojej činnosti na človeka i celú spoločnosť veľkou neznámou.

Chápeme, že ak by sme chceli mať limitovanú AGI „pod kontrolou“, musíme zabezpečiť nielen to, že konečné ciele jej činnosti stanovujeme my, ale dokážeme i ustrážiť spôsob, ako ich chápe a dosahuje, v akom rozsahu a aké parciálne

636 Por. RUSSELL, *Human Compatible*, s. 169.

637 Koncept limitovanej AGI sme uviedli v kapitole 5.3.

prostriedky volí i aké inštrumentálne ciele si stanoví.

Musíme mať na pamäti, že na rozdiel od ANI, pri ktorej je človek podstatným prvkom návrhu a konfigurácie systému i zadávateľom vstupných dát, pri všeobecnej umelej inteligencii sa budeme stretávať so systémami, pri ktorých ani tieto technologické aspekty nebudeme mať pod kontrolou.

Triezvo musíme vnímať, že konečné ciele, ktoré limitovanej AGI budeme predkladať, budú trpieť ľudskou nedokonalosťou a v prípade, že ich strojová inteligencia nesprávne pochopí, môžu sa bytostne obrátiť proti nám. Brániť sa však dokážeme oveľa menej, než u ANI, keďže nebudeme schopní pokryť všetky možné situácie a reakcie pokročilej umelej inteligencie tak, aby sme ochránili ľudské preferencie a človeka.

Nemôžeme očakávať model správania limitovanej AGI analogický ľudskému modelu, ťažko dokážeme zabezpečiť aspekty zdravého rozumu porovnateľné s ľudským náprotivkom a budeme mať problémy vytvoriť ucelenú teóriu strojovej mysle.

Pri pokuse o relektúru riešenia etických výziev, ktoré sme pre ANI uvádzali v 4. kapitole, si uvedomujeme, že **i limitovanú AGI vnímame ako jeden z moderných areopágov, ktorý sa má stať priestorom uskutočňovania Božieho zámeru s človekom a jemu zvereným svetom** – priestorom, do ktorého treba s odvahou vstúpiť a vo svetle evanjelia i v tejto oblasti prispievať k nekončiacemu úsiliu budovať spravodlivý svet, chrániť dôstojnosť každého človeka a rozvíjať civilizáciu lásky.[638]

Avšak diapazón limitovanej AGI pravdepodobne bude priestorom, v ktorom oveľa častejšie zazneje varovný hlas a imperatív zápasu o skutočné dobro človeka a ľudskej civilizácie – napr. v kontexte reálnej transhumanistickej adaptácie AI a extrémnych rizík, ktoré by mohlo zneužitie, resp. nesprávne fungovanie všeobecnej umelej inteligencie predstavovať.

Pre skutočné a úspešné riešenie etických problémov a výziev technológií slabej umelej inteligencie sme zdôrazňovali potrebu interdisciplinárneho rámca, v rámci ktorého by sme mali byť dostatočne oboznámení aj s technologickou stránkou týchto systémov a aj s psychologickými, sociologickými i právnymi aspektmi ich nasadenia.[639] **Pre riešenie výziev limitovanej AGI však interdisciplinárny rámec je a bude nielen primerane potrebný, ale nutný a naviac rozšírený o ďalšie vedné odbory**. Ďalej predpokladáme,

638 Kapitola 4.1.

639 Kapitola 4.2.

že viaceré problémy nebude ani možné riešiť primárne v technologickej rovine, ale skôr v oblastiach, v ktorých sa bude snúbiť filozofia, sociológia, psychológia, kybernetika i ďalšie odbory schopné zahrnúť aj neuro a bio technológie. Samotný vedný odbor umelej inteligencie bude musieť prejsť evolučným (možno revolučným) vývojom, ktorý celý odbor posunie na kvalitatívne vyššiu úroveň (nielen technologicky, ale aj interdisciplinárne).

Základný princíp, ktorý sme stanovili – umelá inteligencia zameraná na dobro človeka[640] – má univerzálnu platnosť pre akékoľvek systémy umelej inteligencie a naplno sa vzťahuje i na technológie všeobecnej umelej inteligencie. Napriek tomu, že ide o princíp, ktorý požíva univerzálny konsenzus, radi by sme pripomenuli, že i v oblasti limitovanej AGI treba vyžadovať jeho aplikáciu v plnej šírke tak, ako sme uvádzali v kapitole 4.3.1.

Návrh riešenia etických požiadaviek na dôveryhodné a na dobro človeka orientované systémy AI, ktorý sme predstavili v 4. kapitole a sumarizovali na obr. č. 12, považujeme za opodstatnený, adekvátny a potrebný i v prípade všeobecnej umelej inteligencie... S jedným podstatným rozdielom: na základe prekročenia koncepčných zlomov by niektoré požiadavky boli takmer automaticky vyriešené[641], riešenie mnohých iných principiálne odlišné[642] a pribudli by požiadavky úplne nové[643].

To, čo by mohlo byť plne v našich rukách, je principiálny postoj spoločnosti k pokročilej umelej inteligencii, pretavený do legislatívnych rámcov a pravidiel pre základný i aplikovaný výskum a adekvátne úsilie venované edukácii i osvete spoločnosti.[644]

Čo považujeme za veľkú neznámu a hlavnú výzvu pre splnenie etických požiadaviek na dôveryhodné a na dobro človeka orientované systémy AGI, je adekvátne riešenie problémov, ktoré sme uvádzali v kapitole 5.4.

Problematika AGI a jej rizík bola imanentne prítomná už od pionierskych čias vývoja umelej inteligencie. Konkrétne kontúry a reálny obsah však dostáva až v poslednom

640 *Human-centered artificial intelligence*, resp. *beneficial artificial intelligence*. Kapitola 4.3.1.

641 Napr. niektoré aspekty robustnosti.

642 Napr. viaceré aspekty bezpečnosti.

643 Medzi ne by napr. patrilo riešenie problémov, ktoré sme uvádzali v kapitole 5.4.

644 Por. Kapitola 4.3.2. a 4.3.4.

období, keď sa javí, že striedanie ročných období vývoja systémov AI sa preklápa do permanentného rozvoja inteligentných strojov.[645]

Hlbšie skúmanie problematiky všeobecnej umelej inteligencie tak aj schladilo prvotné predstavy jednoduchých riešení, medzi ktoré patrili návrhy ako:[646]

- akonáhle budeme mať k dispozícii stroj s vysokým stupňom inteligencie, zistíme, ako ho ovládať
- aj keď vytvoríme systém AI, ktorý bude pre nás nepochopiteľný, problémy dokážeme riešiť ak:
  - systém AI bude emulovať celý mozog, resp. bude vytvorený ako elektronická kópia ľudského mozgu[647]
  - systém AI vznikne na základe metód založených na simulovanej evolúcii programov[648]

Či už tieto návrhy alebo sofistikované riešenia, ktoré používame v rámci vývoja a eliminácie rizikových faktorov symbolických i subsymbolických technológií slabej a úzkej umelej inteligencie, nebudú v prípade AGI fungovať.

Jednou z vecí, ktoré preto musíme poopraviť, je naše tvrdenie zo začiatku tejto kapitoly o cieľoch.[649] Principiálne je správne, no prakticky sa k riešeniu musí pristupovať inak, keďže pri AGI v konečnom dôsledku nemusíme byť ani schopní nejaké inštrumentálne ciele stanovovať, chápať dôsledky a správnosť stanovených konečných cieľov a byť aj schopní kontrolovať, či umelou inteligenciou zvolené parciálne ciele, resp. procesy a zdroje použité na ich dosiahnutie nie sú

645 Eventuálny postupný prerod striedania ročných období AI do prakticky kontinuálneho rastu a rozvoja sme diskutovali v kapitole 1.8.

646 Por. RUSSELL, *Human Compatible*, s. 171.

647 SANDBERG, A. *Whole brain emulation: A roadmap*. Future of Humanity Institute. Oxford University, 2008.

648 KOZA, J. *Genetic Programming: On the Programming of Computers by Means of Natural Selection*. MIT Press, 1992.

649 „Chápeme, že ak by sme chceli mať limitovanú AGI 'pod kontrolou', musíme zabezpečiť nielen to, že konečné ciele jej činnosti stanovujeme my, ale dokážeme i ustrážiť spôsob, ako ich chápe a dosahuje, v akom rozsahu a aké parciálne prostriedky volí i aké inštrumentálne ciele si stanoví.“

problematické, nesprávne alebo nebezpečné.[650]

Základný princíp – umelá inteligencia zameraná na dobro človeka – sa takto primárne transformuje na výzvu: ako zabezpečiť, aby inteligentné stroje dosahovali skôr ľudské ciele než svoje? A ako to dosiahnuť i v prípade, že by systémy AGI nevedeli, aké sú naše ciele?[651]

5.5.3. Pokus o riešenie – hrubé kontúry obrazu, ktorého detaily stále chýbajú

Prehodnotenie formulácie základného princípu u limitovanej AGI v podstate znamená požiadavku na prehodnotenie štandardného modelu, v rámci ktorého sú systémy ANI vytvárané optimalizačnými strojmi, pracujúcimi na základe nami zadaných cieľov. Bez tohoto základného prehodnotenia nie sme schopní vyvíjať systémy limitovanej AGI, ktoré by skutočne boli zamerané na dobro človeka.

Tiež si treba uvedomiť – a limitovaná AGI nás k tomu priamo zo svojej definície pobáda – že **v prípadnej konfrontácii všeobecnej umelej inteligencie s ľuďmi nemôžeme hovoriť o vedomom postavení sa proti človeku, ale skôr o dôsledkoch nesprávneho použitia jej vysokej inteligencie** v rámci dosahovania stanovených cieľov, resp. s tým súvisiacich činností, čo by samo o sebe mohlo mať katastrofálne dôsledky pre ľudskú civilizáciu.

I keď v súčasnosti na riešení rizík AGI a prehodnotení štandardného modelu pracujú

650 Prof. Russell to vysvetľuje slovami: „Podobne ako v mnohých iných oblastiach, aj v oblasti umelej inteligencie bol prijatý štandardný model: vytvoríme optimalizačné stroje, zadáme im ciele a ony začnú pracovať. Toto fungovalo dobre, pokiaľ boli stroje hlúpe a mali obmedzený rozsah činnosti; ak ste vložili nesprávny cieľ, mali ste dobrú šancu, že budete môcť stroj vypnúť, opraviť problém a skúsiť to znova. Keďže sa však stroje navrhnuté podľa štandardného modelu stávajú inteligentnejšími a ich pôsobenie sa stáva globálnejším, tento prístup sa stáva neudržateľným. Takéto stroje budú sledovať svoje ciele bez ohľadu na to, aké zlé by to bolo; budú odolávať pokusom o ich vypnutie; a budú získavať všetky zdroje, ktoré prispievajú k dosiahnutiu cieľa. Optimálne správanie stroja môže naviac zahŕňať oklamanie ľudí, aby si mysleli, že dali stroju rozumný cieľ, čím by stroj získal dostatok času na dosiahnutie skutočného cieľa, ktorý mu bol daný. A vôbec by to nebolo 'deviantné' alebo 'zlomyseľné' správanie, ktoré si vyžaduje (s)vedomie a slobodnú vôľu; bola by to len súčasť optimálneho plánu stroja na dosiahnutie svojho cieľa.“ RUSSELL, *Human Compatible*, s. 172.

651 Prof. Russell je presvedčený, že vyriešenie týchto otázok by nakoniec malo viesť k strojom, ktoré pre nás nepredstavujú žiadnu hrozbu, bez ohľadu na to, aké sú inteligentné. RUSSELL, *Human Compatible*, s. 173.

špičkové tímy v tých najlepších centrách výskumu umelej inteligencie, určite nebude ľahké nájsť skutočné a definitívne riešenie tejto problematiky.

V rámci tohoto hľadania by sme radi predstavili a diskutovali návrh riešenia prof. Russella, o ktorom si myslíme, že patrí k tomu najlepšiemu, čo súčasný výskum AGI môže ponúknuť.

Na základe doteraz uvedenej analýzy – v našom podaní kapitoly 5.3.-5.5. – **Stuart Russell definuje tri princípy vývoja a tvorby na dobro človeka orientovanej AGI**:[652]

- **jediným cieľom inteligentného stroja je maximalizovať realizáciu ľudských preferencií**
- **inteligentný stroj si na začiatku nie je istý, aké sú tieto ľudské preferencie**
- **základným, definitívnym a konečným zdrojom informácií o ľudských preferenciách je ľudské správanie**

Tieto princípy by nemali byť chápané ako explicitné návody, resp. zákony, ale skôr ako usmernenia a ideové návrhy pre vývoj na dobro človeka zameranej AGI.

Prof. Russell narába s termínom preferencie a jeho rôznorodosťou nasledovne:[653]

- preferencie sú všeobjímajúce – pokrývajú všetko, na čom nám môže záležať, a to ľubovoľne ďaleko do budúcnosti
- stroj sa nesnaží identifikovať alebo prijať jeden ideálny súbor preferencií, ale pochopiť a uspokojiť (v rámci možností) preferencie každého človeka

1. princíp: čisto altruistické stroje

Maximalizácia realizácie ľudských preferencií ako jediný cieľ inteligentného stroja je podmienkou *sine qua non* pre systémy na dobro človeka zameranej AGI.

V rámci tohoto princípu sú systémy AI tak zamerané na dobro človeka, že nepripisujú absolútne žiadnu vnútornú hodnotu vlastnému blahu alebo dokonca vlastnej existencii.[654]

Z hľadiska stanovovania a napĺňania inštrumentálnych i konečných cieľov ide o korektne definovaný princíp, no jeho realizácia naráža na viaceré úskalia.

652 RUSSELL, *Human Compatible*, s. 173.

653 RUSSELL, *Human Compatible*, s. 173.

654 RUSSELL, *Human Compatible*, s. 173.

Pre vývoj systémov AI sú ľudské preferencie a učenie sa z ľudského správania veľmi neisté a ťažko z ľudského správania uchopiteľné skutočnosti, napr.: ako ich vedieť správne pochopiť, implementovať a systémovo s nimi narábať, ako sa stavať k dynamike ich vývoja v rámci ľudského života, ako sa vysporiadať s nekalými preferenciami konkrétnych jednotlivcov a pod. Súčasne povinnosť pochopiť a splniť do čo najväčšej miery preferencie každej osoby prináša otázky výberu medzi preferenciami viacerých osôb, či už ide o ich rôznorodosť alebo o porovnanie z pohľadu etiky a morálnych hodnôt..

V kontexte týchto problémov si myslíme, že v rámci naplnenia 1. princípu ľudské preferencie nestačia, treba ich kombinovať s hodnotami. Uvedomujeme si tiež, že v celospoločenskom diskurze môže ísť o kardinálny problém, keďže sa naprieč ľudskou civilizáciou asi nezhodneme, ako by mali tieto hodnoty vyzerať.[655]

Vzhľadom na nereálnosť priamej implementácie hodnôt v rámci technológií AGI sa proces ich derivácie z ľudských preferencií javí ako vhodná cesta pre vývoj na dobro orientovanej strojovej inteligencie.[656] Avšak bez jasného definovania a ukotvenia hodnotového rámca tak znovu vyvstáva riziko relativizácie hodnôt analogické nielen rizikám následnej regulácie u LAWS,[657] ale naviac prerastajúce do panteónu dobra a zla, z ktorého by mohli systémy AGI – pozorujúc dnešný svet – čerpať.

Ide tak o ďalšiu veľkú výzvu pre vývoj limitovanej AGI – **ako definovať a ukotviť hodnotový rámec, resp. ako ho aplikovať na hodnoty, ktoré inteligentný systém derivuje z ľudských preferencií**.

Taktiež treba poznamenať, že ako pri riešení mnohých iných vedeckých problémov, i pri vývoji AGI sa využívajú určité zjednodušenia a aproximácie, bez ktorých by sme sa

---

655 Sám Russell bol v rámci komunikácie tohoto princípu konfrontovaný s novinárskou poznámkou: „Čo dáva západným, dobre situovaným, bielym 'cisgender' mužom a vedcom, ako je Russell, právo určovať, ako stroj implementuje a rozvíja ľudské hodnoty?“
ELKUS, A. *How to be good: human values to artificial intelligence*. Slate, April 20, 2016.

656 Nielen z dôvodu nereálnosti priamej implementácie, ale i na základe toho, čo bolo doteraz uvedené v kapitolách 5.4. a 5.5., môžeme zabudnúť aj na aplikáciu hodnôt v rámci regulačných rámcov, ktoré sa využívajú u systémov ANI.

657 Fenomén „následnej regulácie“, ktorá smeruje k hodnotovému a morálnemu relativizmu a snaží sa prehodnotiť argumenty o nenahraditeľnosti ľudského svedomia a morálneho úsudku, sme v kontexte smrtiacich autonómnych zbraňových systémov diskutovali v kapitole 2.7.8.

v tomto štádiu výskumu zamotali a nevedeli pokročiť ďalej. Takže vývoj pokračuje...

2. princíp: neistota strojov spoznávajúcich ľudské preferencie

Z analýzy rizík systémov limitovanej AGI chápeme, že inteligentné stroje s jasne definovanými ľudskými preferenciami by mali veľký problém. Boli by nielen náchylné ignorovať ľudské pokusy o opravu či zastavenie ich činnosti, no naviac by sa vo svojej logike dosahovania konečných cieľov mohli postupne viac a viac vzďaľovať od myslenia a zámerov ľudí. Preto vnímanie určitej neistoty ohľadom ľudských preferencií, t. j. cieľa, zámerov a zmyslu svojej činnosti by limitovanú AGI viedlo k relativizácii svojej činnosti a tým aj k umožneniu svojho vypnutia, limitovania zdrojov, či akéhokoľvek iného obmedzenia. Stroje by tak zostali závislé na ľudskej kooperácii a spätnej väzbe ohľadom preferencií a smerovania ich činnosti.

Princíp „inteligentný stroj si na začiatku nie je istý, aké sú ľudské preferencie“ je tak podľa prof. Russella kľúčom k úspešnému vytvoreniu na dobro človeka zameraného systému AGI.[658]

Ide teda o rozšírenie princípu neistoty i na preferencie – de facto ciele inteligentných strojov. Je na povážení, že princíp neistoty, ktorý stál pri zrode moderných algoritmov AI,[659] sme až doteraz ignorovali v oblasti stanovovania cieľov, funkcie užitočnosti, resp. odmeňovania, ceny a strát, atď. Vyzerá to na principiálny problém pre ďalší pokrok v teórii riadenia, štatistike, strojovom učení a pod.

Ignorovanie princípu neistoty nie je ani tak o technickej náročnosti, ako skôr o našom myšlienkovom rámci: podľa vzoru ľudskej inteligencie (resp. jej symbolickej úrovne uvažovania a myslenia), ktorá si stanovuje ciele, kritériá a postupy na ich dosiahnutie, sme analogicky vytvárali inteligenciu strojovú.[660]

I keď sa dá povedať, že 2. princíp môže úspešne eliminovať viaceré podstatné riziká, ktoré sme v kapitole 5.4. uvádzali (napr. problémy s inštrumentálnymi cieľmi „strachu“, t.j. sebaobrany a „chamtivosti“, t.j. alokácie prostriedkov), nebude ľahké sa vymaniť z pasce

---

658 RUSSELL, *Human Compatible*, s. 175.

659 Problematiku neistoty sme diskutovali v kapitole 1.4. v rámci porovnávania symbolických a subsymbolických systémov.

660 Por. RUSSELL, *Human Compatible*, s. 176.
Dokonca i pri tvorbe subsymbolických systémov, ktoré stavajú na princípe neistoty, sme sa v oblasti stanovovania cieľov tejto analógie nezbavili.

hodnotovej relativizácie v rámci spoznávania ľudských preferencií.

3. princíp: proces učenia sa predikcie ľudských preferencií

Proces učenia sa a spoznávania ľudských preferencií je striktne ohraničený požiadavkou, aby základným, definitívnym a konečným zdrojom informácií o týchto preferenciách bolo ľudské správanie. Sledujeme tak dva zámery.

Ponajprv ide o vyjasnenie si pojmu a zdroja ľudských preferencií, ktoré je možné spoznávať a chápať len na základe ľudských rozhodnutí, resp. volieb a výberov. Ľudské rozhodnutia sú prejavom ľudských preferencií a pre inteligentný stroj zároveň filtrom preferencií, ktoré sú pre jeho činnosť relevantné.

Sekundárnym zámerom je snaha umožniť inteligentným strojom prehlbovať svoje poznanie a chápanie ľudských preferencií (optimalizovať tento model), a tak zefektívňovať a skvalitňovať svoju činnosť.

I tento princíp sprevádzajú výzvy a riziká, ktoré bude treba seriózne riešiť: ľudia totižto nemusia vždy konať racionálne a vzťah medzi ľudskými preferenciami a ľudskými voľbami nemusí byť presný. To znovu evokuje otázku poznania ľudského modelu správania, bez čoho asi nebude možné v plnej miere realizovať učenie sa a spoznávanie preferencií na základe ľudských volieb.

Znovu opakujúcim problémom sa i v prípade 3. princípu javí disproporcia medzi spoznávaním ľudských preferencií z ľudských volieb a schopnosťou takto naučené preferencie zasadiť do adekvátneho hodnotového rámca. Možnosť disproporcie sa nápadne podobá Bostromovej *téze o nezávislosti*[661] – naučené preferencie a hodnotový rámec nemusia korelovať.

Chápanie týchto troch princípov vývoja a tvorby na dobro človeka orientovanej všeobecnej umelej inteligencie v rovine usmernení a ideových návrhov pre vývoj systémov limitovanej AGI môže zvádzať k ich bagatelizovaniu na úroveň rámcových a orientačných technologických doporučení. Avšak **na základe doteraz uvedenej rozpravy musíme akcentovať potrebu vybudovania pevných základov budúcich technológií AGI, pričom práve tieto princípy môžu tvoriť seriózny ideový základ**.

661 Ide o alúziu na Bostromovu *tézu o nezávislosti* (the orthogonality thesis), ktorú uvádzame v kapitole 5.5.1.: „Inteligencia a konečné ciele sú nezávislé: viac-menej akúkoľvek úroveň inteligencie možno v zásade kombinovať s viac-menej akýmkoľvek cieľom.“

Riziká a problémy, ktoré sme uvádzali v kapitole 5.4., sú dôrazným varovaním, že dobré úmysly, vzdelávacie a osvetové iniciatívy, priemyselné kódexy správania, právne predpisy a ekonomické stimuly konať správne vôbec nemusia nestačiť.[662]

Preto musíme hľadať presné definície a exaktné matematické dôkazy, ktoré by poskytli nespochybniteľné záruky pre striktné zameranie systémov AGI na dobro človeka. **Naviac si musíme byť v čo najväčšej miere istí, že tieto záruky sú skutočne tým, čo chceme, a že predpoklady, ktoré sú súčasťou našich riešení, sú skutočne pravdivé**.[663]

I napriek výhradám a etickým podnetom, ktoré sme v rámci diskusie jednotlivých princípov uviedli, pokladáme návrh prof. Russella za dôležitý vklad do riešenia vývoja limitovanej AGI a eliminácie rizík, ktoré s touto pokročilou inteligenciou budú prichádzať. **Russellove princípy tak majú potenciál dať základ serióznemu dizajnu limitovanej AGI, prekračujúcej koncepčné limity na ceste k skutočnej inteligencii**.

Zároveň však upozorňujeme na riziko relativizácie hodnôt ako dôsledok, resp. problém nielen navrhnutého riešenia, ale akéhokoľvek projektu vývoja všeobecnej umelej inteligencie, čo vyplýva už z podstaty AGI.

Je nám jasné, že naša snaha zakomponovať primeraný hodnotový rámec do vývoja limitovanej AGI môže ísť proti zaužívanému naratívu falošného chápania ľudskej slobody postmoderného človeka, no **bez implementácie hodnotového rámca by sme síce mohli odvrátiť nebezpečenstvo silnej singularity ASI – superinteligencie, ktorá prevýši náš intelekt, no takmer určite by sme padli do oveľa sofistikovanejšej pasce slabej singularity – umelej inteligencie, ktorá definitívnym spôsobom ovládne a prekoná naše slabosti**.[664]

V pomyselnom zápase s umelou inteligenciou by nás nedeklasovala nejaká super inteligencia, ktorá sa vedome postaví proti nám, ale nakoniec by sme si strelili gól do vlastnej bránky sami – katastrofálnymi dôsledkami nesprávneho dizajnu, použitia a činnosti limitovanej AGI.

---

662 „Keď je v stávke budúcnosť ľudstva, nádej a dobré úmysly, vzdelávacie iniciatívy a priemyselné kódexy správania, právne predpisy a ekonomické stimuly pre správne konanie nestačia.“ RUSSELL, *Human Compatible*, s. 184.

663 Por. RUSSELL, *Human Compatible*, s. 184.

664 Dva pohľady na singularitu v oblasti umelej inteligencie sme schematicky vyjadrili na obr. č. 10.

Zhrnutie na záver

Rozprávanie o stvorení v knihe Genezis, s jej potvrdením poriadku a dobroty všetkého stvorenia ako prejavu jediného a milujúceho Boha, vyjadruje základný vzťah a prvotnú zmluvu[665], ktorú Boh uzavrel s ľudstvom.

Meditácia tohoto pohľadu, spojená s gréckou metafyzickou špekuláciou o bytí a kauzalite, jednote a rozdielnosti, pohybe, zmene a číslach, umožňovala človeku chápať výskum prírodných vied ako „čítanie knihy stvorenia“.[666] A v kontexte dejinného vývoja s príchodom scholastickej metódy[667] a takých princípov, ako je napr. Occamova britva, sa ľudstvo usilovalo vziať do ruky rydlo, neskôr brko s kalamárom a nakoniec moderné pero, aby do tejto knihy prírody aj niečo doplnilo.

Ak však máme v tejto analógii pokračovať ďalej, rozvoj umelej inteligencie spojený v rámci štvrtej priemyselnej revolúcie s napredovaním v robotike a biotechnológiách možno prirovnať ku Gutenbergovmu stroju, ktorý znamenal prelom v tom, čo a ako dokázal človek zapísať – teda ako a v čom chceme knihu stvorenia dopĺňať alebo prepísať. Je tak na nás, či sa dokážeme stále vracať k *počiatkom* (~Genezis) a mať dobre premeditované to, čo nám Boh zjavil, aby sme tak „Gutenbergov vynález umelej inteligencie“ správne použili a tento svet skutočne rozvíjali. Veď návod „ako na to“ sme v Ježišovi Kristovi a v evanjeliu dostali a treba ho len rozmeniť na drobné, aby sme ho aj v oblasti umelej inteligencie napĺňali.

V publikácii sme sa primárne venovali problematike etických výziev a morálnych aspektov súčasných systémov umelej inteligencie, ktorú poznáme pod spoločným označením slabá umelá inteligencia (ANI). Tento termín zahŕňa úzko špecializované systémy umelej inteligencie (narrow AI), ktoré sú optimalizované na zvládnutie konkrétnej

665 Por. ŠANTAVÝ, P. Dejiny spásy: zmluvy Starého zákona a Nová zmluva [diplomová práca]. [on-line]. Bratislava: RKCMBF UK, 2000. [cit. 6. novembra 2021]. Dostupné na internete: <https://peter.santavy.cloud/data/uploads/docs/dejiny_spasy-zmluvy_sz_a_nz.pdf>

666 Por. SMITH, R. *Why No Science?* [on-line]. [cit. 6. novembra 2021]. Dostupné na internete: <https://www.thecatholicthing.org/2021/11/02/why-no-science/>

667 *Scholastic Method* [on-line]. [cit. 6. novembra 2021]. Dostupné na internete: <https://www.encyclopedia.com/religion/encyclopedias-almanacs-transcripts-and-maps/scholastic-method>

úlohy, resp. množiny úloh. Ide súčasne o systémy slabej umelej inteligencie (weak AI), ktoré vykazujú inteligentné správanie na základe modelov a aplikovaných metód i tréningových dát. Hovoríme teda o systémoch, ktoré sú zamerané na riešenie konkrétnych úloh a sú závislé na ľudskom vstupe a konfigurácii.

Jedným z hlavných prínosov tejto publikácie je ponúknutý kvalitný a primerane hlboký interdisciplinárny rámec, bez ktorého nie je možné úspešne realizovať skutočné riešenie etických problémov a výziev technológií umelej inteligencie. Ide o rámec, v ktorom sme dostatočne oboznámení aj s technologickou stránkou týchto systémov a psychologickými, sociologickými i právnymi aspektmi ich nasadenia.

V prvej kapitole sme preto ponúkli potrebný náhľad do problematiky umelej inteligencie, sumarizujúc jej základné vlastnosti, delenie a metódy. Osobitne sme sa venovali algoritmom inšpirovaným činnosťou mozgu, keďže ony sú základom drvivej väčšiny pokročilých systémov AI. Poukázali sme na základné problémy niekoľkých dekád vývoja systémov AI i na masívne zavádzanie týchto technológií v súčasnosti, pričom nezostal opomenutý ani futuristický presah fenoménu umelej inteligencie.

Druhá kapitola predstavuje základ pre **ďalší dôležitý prínos tohto diela – identifikáciu, pomenovanie, analýzu a pochopenie rizík spojených s technológiami umelej inteligencie v celej možnej šírke spektra ich nasadenia.**

Pre reálne uchopenie problematiky etiky týchto technológií považujeme široko spektrálne uchopenie a komplexné pochopenie limitov a rizík súčasných systémov AI za podstatné. Preto sme sa v druhej kapitole pomerne obšírne venovali hlavným rizikovým faktorom a zraniteľnostiam algoritmov súčasných systémov AI, zaoberali sme sa bezpečnosťou procesov založených na týchto technológiách, vysvetľovali problematiku kybernetickej bezpečnosti ako integrálnej súčasti zabezpečenia funkčnosti systémov AI a neopomenuli sme ani riziká vyplývajúce z technologickej komplexnosti a potrebného infraštruktúrneho zázemia pre spoľahlivú činnosť v reálnom svete.

Analýza uvedených limitov a rizík bola základom pre rozoberanie negatívnych dôsledkov využívania technológií umelej inteligencie v spoločnosti, zahŕňajúc celé spektrum ich nasadenia od sociálnych sietí až po problematiku autonómnych vozidiel a rozširujúc tak interdisciplinárny rámec o psychologické a sociologické dôsledky ich využitia. Osobitne sme sa venovali oblasti dohľadových systémov, technológií využívaných v spravodajských službách a v rámci algoritmického riadenia štátu, keďže z pohľadu etických výziev ide

o veľmi špecifické pole vývoja i nasadenia systémov umelej inteligencie. Azda najproblematickejšou oblasťou je adaptácia technológií AI vo vojenskej sfére – od využitia v armádnych spravodajských službách, cez modelovanie a simulácie technológií a procesov, virtualizačné nástroje a výcvik, až po vývoj a nasadenie smrtiacich autonómnych zbraňových systémov a kybernetických zbraní. Armádne využitie systémov AI v sebe obnáša celé spektrum otázok s potenciálom prevýšiť všetky ostatné etické dilemy a výzvy, preto sme tejto oblasti venovali osobitný priestor.

V prvej časti tretej kapitoly sme **ako ďalší osobitný prínos sumarizovali naše etické postrehy a závery, ku ktorým sme dospeli v rámci analýzy limitov a rizík súčasných systémov AI**:

- súčasné systémy AI sprevádzajú oprávnené a vážne obavy z toho, že ak nechápeme ako tieto systémy pracujú (prakticky každý sofistikovaný systém umelej inteligencie sa javí ako black box), nemôžeme im reálne dôverovať a ťažko dokážeme predpovedať okolnosti, za ktorých tieto systémy zlyhajú.
- problém s verifikovateľnosťou postupov a výsledkov činnosti systémov umelej inteligencie sa tak premieta i do oblasti implementácie etických pravidiel, či už ide o spôsob, ako hodnotiť a kontrolovať činnosť systémov AI, alebo ide priamo o spôsob implementácie regulácií technológiami AI.
- súčasné systémy umelej inteligencie trpia rôznymi zraniteľnosťami. Musíme mať neustále na pamäti, že tieto systémy môžu zlyhávať najrozličnejšími a často neočakávanými spôsobmi. Systémy umelej inteligencie robia chyby, ktoré sú diametrálne odlišné od ľudských chýb a zlyhaní. Preto môžu byť prekvapivé, neočakávané a nebezpečné.
- s týmito zlyhaniami a obmedzeniami treba rátať aj v oblasti snahy o technologickú implementáciu etických noriem a mantinelov. Nemusí to byť vôbec triviálne a v niektorých scenároch nasadenia ani uskutočniteľné.
- vzhľadom na rozličné typy útokov zameraných na procesy umelej inteligencie sa etické výzvy neviažu len na návrh a realizáciu systémov AI, no rozširujú sa aj o oblasť etiky použitia a z toho prameniace riziká zneužitia.
- systémy umelej inteligencie sú vystavené aj problémom v oblasti kybernetickej bezpečnosti, keďže mnohé riziká atakujúce bezpečnosť procesov umelej

inteligencie, zneužívajúce nedostatky dizajnu alebo amplifikujúce dôsledky rizikových faktorov, ako vektor útoku používajú zraniteľnosti v oblasti kybernetickej bezpečnosti. Kybernetická bezpečnosť je proces, ktorý má svoju dynamiku a vyjadruje skutočnosť, že ani v oblasti systémov umelej inteligencie neexistuje dokonale bezpečný a spoľahlivý systém.

- vývoj a nasadenie sofistikovaných systémov AI poukazuje na ich veľkú komplexnosť, ktorá je nielen rizikovým faktorom bezpečnosti a stability fungovania systémov, ale mnohokrát i neľahkou výzvou pre úspešnú a jasnú implementáciu regulácií a etických pravidiel.
- úspešné nasadenie mnohých moderných systémov AI vyžaduje prísun extrémneho množstva dát z reálneho sveta a priamo z ľudského prostredia. Pri veľkých systémoch zhromažďujúcich denno denne extrémne množstvo dát je veľmi ťažké zaručiť etiku spracúvania týchto údajov a vylúčiť riziko ich zneužitia.
- systémy umelej inteligencie sa s veľkou mierou istoty stávajú schopnými vytvárať veľmi presné psychologické profily, odhaľovať akékoľvek väzby a mnohé osobné informácie, predikovať a ovplyvňovať naše konanie. Tieto systémy však mnohokrát nie sú predmetom regulácie, resp. verejnej kontroly.
- neuvedomujúc si, ako je naša myseľ a psychika zraniteľná, existuje riziko postupného prechodu od technologického prostredia založeného na systémoch AI k prostrediu založenému na závislosti a manipulácii.
- míľnikom, ktorého by sme sa mali obávať, teda nie je budúca technologická singularita v oblasti umelej inteligencie, v ktorej AI prevýši náš intelekt, ale oveľa skôr moment, keď technológia ovládne a prekoná naše slabosti.
- v kontexte slabej umelej inteligencie (ANI) ešte stále konečnú voľbu cieľov nerobia stroje, ale človek, takže stále je v ľudskej moci tieto dôsledky ovplyvniť a zmeniť. Preto musia existovať etické pravidlá a regulácie, ktoré by dokázali chrániť jednotlivca i celú spoločnosť voči rizikám sofistikovanej práce systémov AI v rámci sociálnych sietí, systémov riadenia spoločnosti, resp. analogických platforiem.
- viaceré aplikácie systémov umelej inteligencie v reálnom svete budú musieť v zlomkoch sekúnd riešiť rôzne kritické situácie – a vedieť ich vyriešiť eticky. Otázkou je, do akej miery môžu byť etické princípy a tomu zodpovedajúce právne

regulácie aj technicky úspešne implementovateľné.

- vo viacerých oblastiach nasadenia bude treba vyriešiť rozdelenie zodpovednosti medzi človekom a riadiacim systémom AI a s tým súvisiace legislatívne požiadavky pre reálnu prevádzku týchto systémov.
- musíme sa seriózne zamyslieť, ako stanoviť akceptovanú mieru neistoty v oblasti spoľahlivosti a robustnosti systémov umelej inteligencie.
- fenomén digitálneho rozdelenia (digital divide), ktorý sa s neustálym rozvojom informačných technológií a prechodom k informačnej spoločnosti stáva reálnym problémom, môže byť vďaka necitlivému nasadeniu technológií umelej inteligencie oproti súčasnosti ešte znásobený.
- v rámci algoritmického riadenia sa zvyšuje riziko zneužitia systémov AI na efektívnu kategorizáciu a obmedzovanie ľudských práv občanov. Aké sú možnosti nastavenia etických mantinelov a reálne dodržiavaných právnych rámcov v oblasti riadenia štátu, spravodajstva a dohľadu?
- trio technológií umelej inteligencie – sledovanie a dohľadové systémy spojené s analytickými nástrojmi a schopnosťou vytvárať modely predikujúce ľudské konanie či vývoj situácie – tvorí v súčasnosti skutočne silnú technologickú výbavu (nielen) spravodajských služieb. Ako spoločnosť musíme vedieť zabezpečiť striktnú zákonnosť i dôsledný dohľad demokraticky zvolených zástupcov a spoločnosti pri nasadení systémov AI v rámci spravodajských služieb, dohľadových systémov a všetkých foriem i stupňov algoritmického riadenia spoločnosti.
- pokročilé modelovanie technológií, simulácia priebehu vojenských operácií i priebehu častí konfliktu a silný virtualizačný efekt môžu viesť k morálnemu znecitliveniu obsluhujúceho personálu a velenia. Dôsledkom môže byť strata vnímavosti pre ohrozenie ľudských životov a reálnych hodnôt pri skutočnom bojovom nasadení.
- existuje silný tlak na využívanie autonómnych zbraňových systémov, ktoré sú vďaka technológiám umelej inteligencie vo viacerých scenároch nasadenia lepšie ako ľudia. Ako dokážeme zabezpečiť, že smrtiace autonómne zbraňové systémy (LAWs), ktoré sú zámerne dizajnované tak, aby dokázali efektívne pracovať aj pri strate spojenia, budú vedieť správne použiť svoju smrtiacu silu a rešpektovať

požadované mantinely, ak je nám známe, že systémy AI robia mnohokrát iné chyby ako ľudia a nevieme ich vytrénovať tak, aby vedeli adekvátne odpovedať na všetky reálne situácie? Uvedený problém ešte narastá pri skupinovom, resp. kooperatívnom riadení autonómnych zbraňových systémov.

- útočné kybernetické systémy využívajúce súčasné technológie umelej inteligencie považujeme za veľmi rizikové a nevhodné (nielen) pre vojenské nasadenie. Keďže nemáme reálne odpovede ohľadom eliminácie ich rizík a nasadenia, stupňuje sa apel proti akémukoľvek nasadeniu útočných autonómnych zbraní, ak sú mimo zmysluplnej ľudskej kontroly.
- v súčasnosti nemáme vedecké dôkazy o schopnosti automatizovaných systémov disponovať funkciami potrebnými na presnú identifikáciu cieľa, situačné povedomie alebo rozhodnutia týkajúce sa primeraného použitia sily. LAWs tak môžu spôsobiť vysokú mieru vedľajších škôd a preto sa rozhodnutia o použití hrubej sily nesmú delegovať na stroje.
- schopnosti umelou inteligenciou poháňaných autonómnych zbraňových systémov a kybernetických zbraní i napriek všetkým rizikám vedú k zvyšujúcim sa tlakom na financovanie a zavádzanie útočných kybernetických zbraní. Je veľkou etickou a regulačnou výzvou, ako tieto tendencie zastaviť.
- problematika obmedzenia týchto útočných systémov je skomplikovaná aj reálnym stieraním hraníc medzi obranným a útočným nasadením takmer vo všetkých oblastiach vojenského využitia technológií umelej inteligencie.
- nutnou podmienkou prevádzky ľubovoľného systému AI, ktorý môže predstavovať riziko pre akúkoľvek ľudskú osobu, je schopnosť a možnosť človeka prebrať kedykoľvek kontrolu nad týmto systémom, resp. právo a možnosť verifikovať a prehodnotiť výsledky jeho činnosti.
- limity, regulácia a obmedzenia LAWs by mali predstavovať etický rámec stanovený na základe morálnych hodnôt ľudskej spoločnosti, nie na základe relativistickej tzv. „následnej regulácie“, ktorý môže viesť k nepredvídateľným morálnym dopadom.
- na základe doterajších skúseností v oblasti limitov a rizík súčasných systémov AI je otázne, či v celom spektre technológií a algoritmov umelej inteligencie bude možné pevný etický rámec dodržať. Vieme si to predstaviť pri cielenom – čo môže

znamenať aj limitujúcom a tým aj znevýhodňujúcom – dizajne systémov AI s dôrazom na dodržiavanie navrhnutého rámca (ethics by design), avšak pri súčasnej snahe o čo najširšie adoptovanie umelej inteligencie v armáde a získanie konkurenčnej výhody sme v tomto smere pomerne skeptickí.

- v kontexte rastúcej dôvery v systémy umelej inteligencie a vnímajúc problémy, s ktorými sa nasadenie týchto systémov potýka i oblastí, ktorých sa ich využitie dotýka, môžeme a musíme stanoviť podmienky, bez splnenia ktorých by nasadenie systémov AI do reálneho sveta, v ktorom interagujú s človekom a vplývajú na spoločnosť, nemalo byť umožnené.

Uvedený sumár problémov technológií umelej inteligencie s priamym či nepriamym dopadom na človeka a spoločnosť sa stal základom pre naše návrhy riešenia etických problémov a stanovenie všeobecných i špecifických etických zásad, ktoré predkladáme v štvrtej kapitole.

Interdisciplinárny rámec sme v rámci tretej kapitoly rozšírili analýzou súčasných aktivít na poli etiky umelej inteligencie smerujúc k potrebným reguláciám na zabezpečenie etického rámca využívania týchto technológií. Osobitne sme sa venovali európskemu Aktu o umelej inteligencii, ktorý považujeme síce za náročnú, avšak v súčasnosti asi najprepracovanejšiu a najkomplexnejšiu (pripravovanú) reguláciu v oblasti umelej inteligencie s potenciálom ovplyvniť etiku využívania systémov AI vo veľkej časti sveta. Tretia kapitola bola zavŕšená analýzou súčasných aktivít Cirkvi na poli etiky umelej inteligencie, pričom sme sa osobitne zaoberali závermi rímskej konferencie renAIssance 2020, známej aj pod názvom Rome Call for Ethics.

Problematika etických výziev a morálnych aspektov súčasných systémov slabej umelej inteligencie (ANI) bola zavŕšená naším návrhom riešenia etických problémov AI, ktoré sme predkladali v štvrtej kapitole. Ponajprv išlo o vyjadrenie a naše chápanie základného, a to pozitívneho postoja k fenoménu umelej inteligencie vo svetle Zjavenia. Osobitne sme akcentovali interdisciplinárny rámec prístupu k problematike etiky AI, bez ktorého skutočné riešenie etických problémov a výziev technológií umelej inteligencie nie je možné úspešne realizovať.

K hlavným prínosom a záverom predkladanej publikácie patrí náš návrh základnej štruktúry etických princípov a zásad, ktorý sme v štvrtej kapitole predstavili:

- rozšírili sme diapazón základného zamerania umelej inteligencie na človeka (human-centered AI, beneficial AI) o kontext kresťanskej antropológie, inklúziu každej ľudskej bytosti bez diskriminácie so zreteľom na dobro ľudstva a spoločnosti v rozšírenej optike starostlivosti o náš spoločný a zdieľaný domov, teda o celý stvorený svet.
- definovali sme a objasnili principiálne požiadavky na dôveryhodné systémy AI, ktoré musia byť legálne, etické a robustné. Vo vedomí, že dokonalý systém umelej inteligencie neexistuje, sme navrhli minimálne legislatívne, etické a technologické požiadavky, resp. hranicu, od ktorej môžeme tieto technológie považovať za dôveryhodné.
- predstavili sme vlastnú sadu všeobecných a univerzálnych etických zásad:
  - pri vývoji, výrobe, nasadení, poskytovaní a používaní systémov umelej inteligencie musí byť zaručená ochrana slobody, dôstojnosti a bezpečia každej ľudskej osoby i celej spoločnosti.
  - technológie umelej inteligencie musia byť plne pod ľudskou kontrolou a ovládateľné človekom.
  - algoritmy i výsledky činnosti systémov AI musia byť človekom pochopiteľné a revidovateľné.
  - akékoľvek nasadenie technológií AI musí byť prospešné pre človeka a spoločnosť.
  - systémy umelej inteligencie nesmú byť nástrojom digitálneho rozdelenia.
  - technológie umelej inteligencie nesmú škodiť nášmu spoločnému domu a mali by prispievať k spoločenskému a environmentálnemu blahobytu.
- nami predstavené etické zásady sme porovnali so zásadami najdôležitejších súčasných aktivít a regulácií v oblasti AI, aby sme tak vyjadrili univerzálnosť a určitú nadčasovosť nášho návrhu.
- naším zámerom bolo navrhnúť a vytvoriť tak univerzálnu a všeobecnú množinu etických zásad, že podľa nej môžeme či už definovať – alebo ešte lepšie – z existujúcich legislatívnych rámcov a etických odporúčaní vyberať konkrétne a jasné odporúčania pre ich aplikáciu v reálnom svete. Takto sme i označili

kombináciu etických odporúčaní vatikánskej konferencie renAIssance 2020 a obsahu európskeho nariadenia Akt o umelej inteligencii za v súčasnosti najvhodnejšie konkrétne a do legislatívneho rámca zasadené odporúčania pre etický vývoj, nasadenie a využívanie súčasných systémov umelej inteligencie.

- tiež sme považovali za dôležité popísať všetky oblasti nutnej implementácie etických noriem, eticko-právnych regulácií a morálnych zásad. Ide o oblasť tvorby systémov AI, poskytovateľov i používateľov týchto systémov a implementácie noriem i obmedzení priamo v systémoch AI. Akcentovali sme viaceré činitele, bez ktorých nie je možné tieto oblasti zasadiť do etického rámca vývoja, nasadenia a vyžívania – ide napr. o edukáciu a osvetu, pravidlá pre základný výskum, nutné podmienky vývoja a pod.

Vzhľadom na dôraz, ktorý sme v druhej kapitole kládli na oblasť pokročilého riadenia štátu, spravodajstva a plošného dohľadu i využívania systémov umelej inteligencie vo vojenskej oblasti, v štvrtej kapitole prinášame i **naše vlastné závery, návrhy regulácií a etické odporúčania v týchto špecifických a dôležitých oblastiach, ktoré považujeme za ďalší osobitný prínos**:

- využitie technológií AI v oblasti pokročilého riadenia štátu, spravodajstva a plošného dohľadu musí byť vzhľadom na svoju povahu, potenciál a riziká pokryté už základnými legislatívnymi mechanizmami a verejným dohľadom demokratickej spoločnosti, ktoré sa týkajú riadenia spoločnosti, ľudských práv a pôsobenia spravodajských služieb vo všeobecnosti.
- oblasť exportu produktov a technológií umelej inteligencie, ktoré môžu byť zneužité v oblasti pokročilého riadenia štátu, spravodajstva a plošného dohľadu, by mala byť predmetom medzinárodnej regulácie s cieľom zamedziť ich vývoz do rizikových krajín.
- technológiami AI poháňané automatické smrtiace zbraňové systémy (LAWs), systémy automatického zameriavania a vyberania cieľov, automatické systémy schopné bez zásahu človeka rozhodnúť o smrtiacej reakcii akéhokoľvek druhu (od útoku dronu až po rozpútanie jadrového konfliktu) musia byť zakázané.
- nutnou podmienkou prevádzky ľubovoľného systému AI, ktorý môže predstavovať riziko pre akúkoľvek ľudskú osobu, je schopnosť a možnosť človeka prebrať

kedykoľvek kontrolu nad týmto systémom, resp. právo a možnosť verifikovať a prehodnotiť výsledky jeho činnosti.

- limity, regulácia a obmedzenia LAWs by mali predstavovať etický rámec stanovený na základe morálnych hodnôt ľudskej spoločnosti, nie na základe relativistickej tzv. „následnej regulácie“.

V závere štvrtej kapitoly **sme predložili niekoľko návrhov pre využitie potenciálu Cirkvi a osobitnú angažovanosť podporujúcu etický prístup k problematike umelej inteligencie v celej jej šírke**. Ide predovšetkým o prínos morálno-etického diskurzu do problematiky etiky AI, ďalej o misiu zjednocovať, usmerňovať a propagovať etické aktivity vo svete, a tiež o dôležitú a neustávajúcu snahu akcentovať a budovať univerzálne bratstvo a sociálne priateľstvo aj v oblasti digitálneho sveta a jeho technológií.

Sekundárnym aspektom nášho úsilia bolo v kontexte súčasných technológií ANI poukázať i na problematiku skutočnej umelej inteligencie (AGI), ktorá by podľa jej protagonistov mala byť dosiahnuteľná prostredníctvom silnej (strong) a všeobecnej (general) AI. Všeobecnej, lebo dokáže zvládnuť akúkoľvek intelektuálnu úlohu a má schopnosť generalizovať, t.j. zovšeobecňovať a prenášať, či adaptovať naučené schopnosti na iné úlohy. Silnej, pretože aj skutočne rozumie tomu, čo rieši a vykonáva.

I keď prakticky pri všetkých témach rozoberaných v prvých štyroch kapitolách sme sa nevyhli aspoň krátkemu pohľadu za horizont – k systémom silnej a všeobecnej umelej inteligencie, až v piatej kapitole sme sa výlučne venovali jej viacerým podstatným problémom, pričom **naše uchopenie problematiky všeobecnej umelej inteligencie taktiež považujeme za prínos, ktorý má svoje opodstatnenie**:

- diskutovali sme teóriu mysle a zdravého rozumu so schopnosťou abstrakcie, analógie, konceptualizácie, simulácie, či chápania zmyslu, identifikujúc tak bariéru chápania zmyslu, venovali sme sa hypotéze stelesnenia a predstavili sme komplexné dôvody, pre ktoré nie sme v súčasnosti schopní vytvoriť silnú a všeobecnú umelú inteligenciu.
- v kontexte nielen klasickej definície, ale i modernistickej snahy o redefiníciu pojmu osoby sme polemizovali s víziou uvedomelej AI a poukázali na dôvody, prečo ani najpokročilejším systémom ANI nemôžeme priradiť štatút osoby.
- na základe kresťanskej antropológie sme vyjadrili výnimočnosť ľudskej bytosti a tým

aj odmietnutie štatútu osoby pre akúkoľvek umelú inteligenciu, vrátane AGI.

- použijúc materialistický uhol pohľadu – odmietajúci duchovnú dušu ako „formu“ tela, zásadne sa podieľajúcu na vedomí a sebauvedomení – sme vyjadrili, prečo sa súhrn veľmi pokročilých ANI nedokáže vyrovnať skutočnej AGI a okrem iného sme poukázali na v súčasnosti neriešiteľnú problematiku implementácie „mechanizmu svedomia“, pokiaľ nebudeme mať vyriešený model správania sa AGI na spôsob teórie mysle a schopnosti zdravého rozumu u človeka.
- osobitne sme akcentovali reálny problém funkčného mechanizmu svedomia, ktoré má referenčný bod mimo seba. U hypotetickej analógie svedomia uvedomelej umelej inteligencie nie sme schopní garantovať, že bude zdieľať rovnaké hodnoty a rovnakým spôsobom ich aj chrániť a zachovávať.
- prezentovali sme naše výhrady voči možnosti vytvoriť všeobecnú umelú inteligenciu, ktorá má vedomie, resp. sebauvedomenie.
- zaviedli sme termín limitovaná AGI, označujúci všeobecnú a silnú umelú inteligenciu, ktorá síce nebude schopná myslieť ako ľudia, t.j. v celej komplexnosti presahujúcej do roviny vedomia a sebauvedomenia, avšak budeme môcť o nej povedať, že myslenie už nesimuluje, ale určitým spôsobom skutočne myslí.
- uviedli sme koncepčné prelomy, ktoré treba zdolať na ceste k limitovanej všeobecnej umelej inteligencii.
- v kontexte očakávaných koncepčných prelomov vo vývoji umelej inteligencie sme rozoberali riziká, ktorými by limitovaná AGI mohla v budúcnosti ohroziť ľudstvo.
- prezentovali sme základný ideový návrh riešenia rizík prameniacich z prekročenia koncepčných prelomov limitovanej AGI a diskutovali sme etické problémy, ktoré na ceste k úspešnému riešeniu týchto rizík bude treba ešte zdolať.

Celkovo toto dielo predkladáme ako určité nóvum, snažiac sa o inovatívny prístup v nazeraní na problematiku umelej inteligencie v kontexte kresťanského svetonázoru a etických dôsledkov limitov a rizík, s ktorými sa táto atraktívna oblasť moderného technologického sveta v súčasnosti potýka. Interdisciplinárne uchopenie fenoménu AI, dôsledná analýza limitov i rizík a z nej prameniace etické závery, návrh základnej štruktúry všeobecných i špecifických etických princípov a zásad, vyjadrenie diskutabilných faktorov i niektorých perspektív všeobecnej umelej inteligencie smerujúcej

k superinteligencii a zdôvodnenie jasného postoja k jej porovnávaniu s ľudskou bytosťou – to všetko tvorí náš vklad do prebiehajúceho celosvetového etického diskurzu, ktorý tým viac naberá na dôležitosti, čím viac sa technológie AI rozvíjajú a jej sofistikované implementácie sa do reálneho života masívne zavádzajú, ovplyvňujúc tak životy miliónov ľudí.

Nástojčivosť tejto témy vynikne, ak si z našej rozpravy o oblastiach implementácie etických princípov (kap. 4.3.4) pripomenieme, že v prípade pokročilých systémov umelej inteligencie – na rozdiel od iných informačných systémov a technológií – prakticky nie je možné na existujúce a už nasadené systémy spoľahlivo aplikovať etické zásady a regulácie prostredníctvom doplnenia technických úprav alebo procesných postupov. Jednoducho povedané, s etikou musíme vývoj týchto systémov sprevádzať a ich realizáciu predchádzať, inak nám – nielen jednotlivcom, ale celej spoločnosti digitálneho veku – tieto technológie prerastú cez hlavu.

Dúfame, že táto publikácia prispeje k rozšíreniu (nielen) etických obzorov v jednej z najdynamickejšie sa rozvíjajúcich oblastí digitálneho sveta a láskavý čitateľ bude zhovievavý k prípadným nepresnostiam a nedostatkom nášho pohľadu na problematiku etických výziev a morálnych aspektov súčasných i budúcich systémov umelej inteligencie.

Poďakovanie autora

Som vďačný Bohu za všetky talenty a možnosti, ktorými ma obdaril i túžbu, ktorá sa mnohorako pretavila aj do oblasti bádania a realizácie projektov v technologickej oblasti.

S láskou si spomínam na nebohých rodičov, ktorí mi boli vzorom a príkladom, pričom ma všemožne podporovali a pomáhali na ceste zhmotnenia túžob do reálnych krokov života.

Zo srdca ďakujem ThDr. Ing. Vladimírovi Thurzovi, PhD., za cenné rady, spätnú väzbu a pripomienky, ktorými ma sprevádzal pri tvorbe tejto publikácie.

Ďakujem mojim blízkym, priateľom i kolegom za podporu, dôveru a trpezlivosť, lebo – aj keď som väčšinu tohto diela písal počas dlhých večerov – boli momenty, v ktorých pociťovali nemalý diskomfort z môjho pisateľského nasadenia.

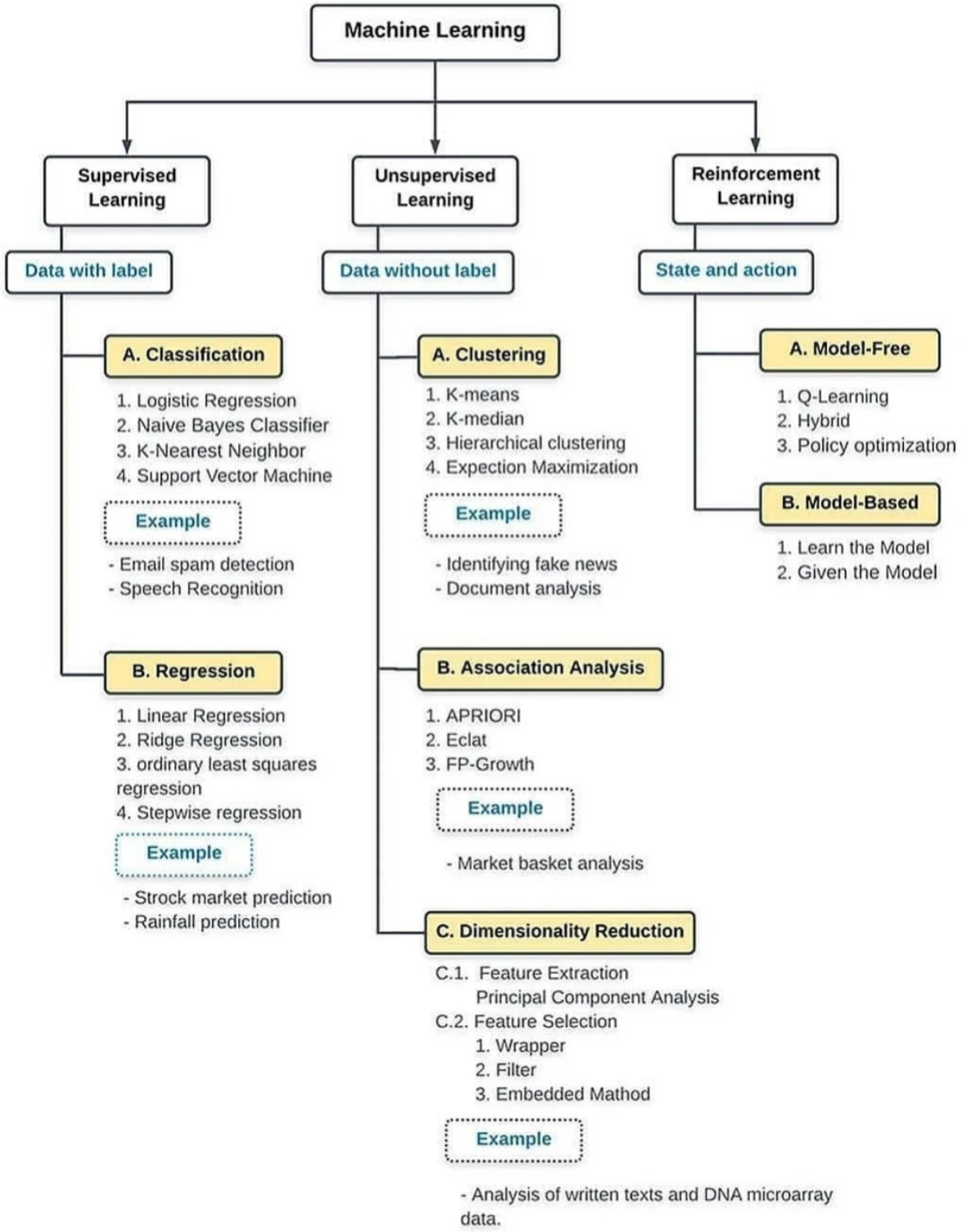
Ďakujem všetkým, ktorí mi boli nápomocní pri vydávaní tejto publikácie, či už išlo o cenné rady pre publikovanie (Mgr. Lujza Rochová, vedúca fakultnej knižnice RKCMBF UK), návrh obálky (pán Miroslav Kulich z Publico), zalomenie a prípravu e-knihy, resp. knihy do tlače (Ing. Matúš Brilla) i samotnú tlač (ExpresTlač).

Pri písaní ma ovplyvnili viaceré osobnosti z oblasti sveta umelej inteligencie:

- futurológ Gerd Leonhard, ktorému vďačím za prvotnú inšpiráciu a chuť písať o tejto problematike,
- špičkoví profesori umelej inteligencie, Melanie Mitchell a Stuart Russell, ktorých vedecká práca mi bola sprievodcom pre zatiahnutie na hlbinu problematiky AI.

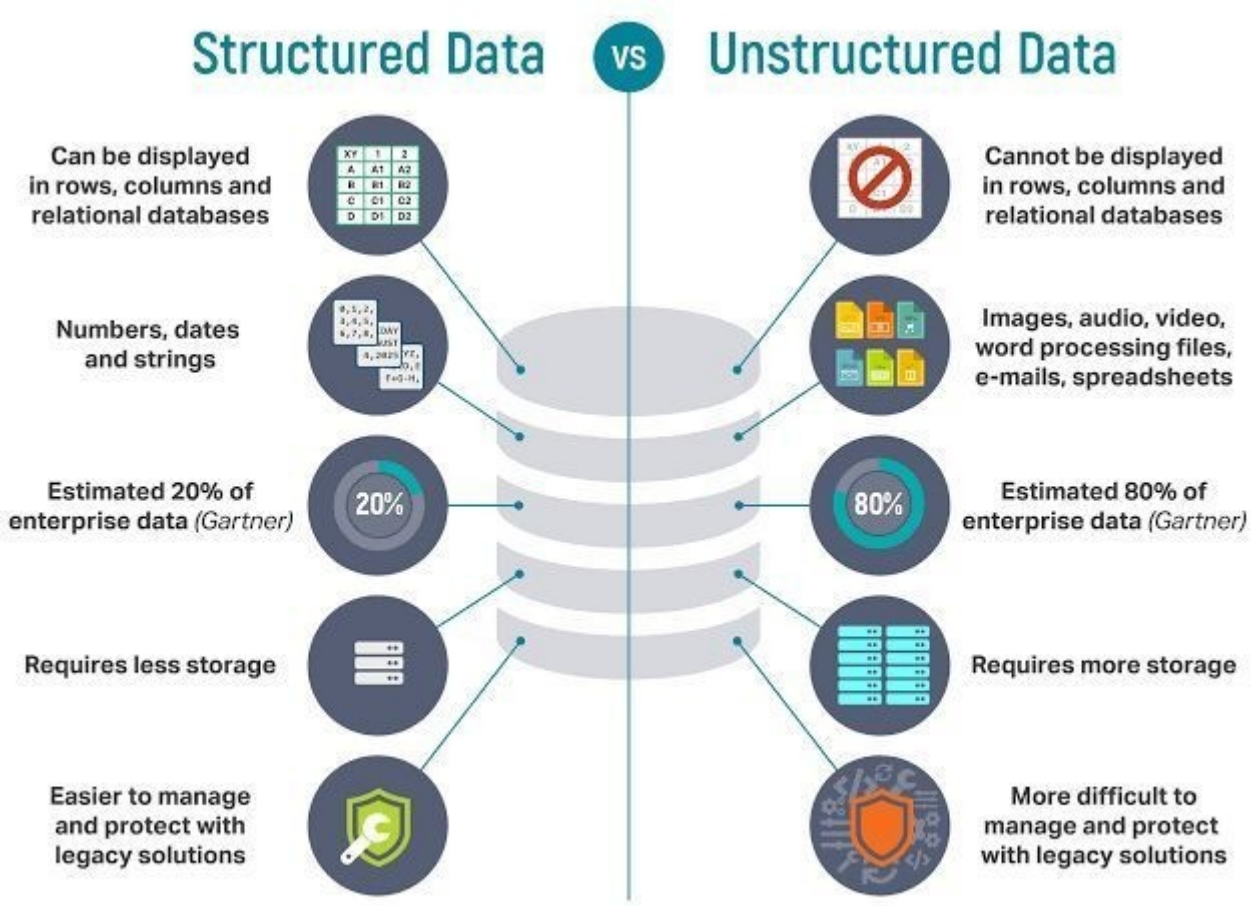
Prílohy

Príloha č. 1 – algoritmy strojového učenia



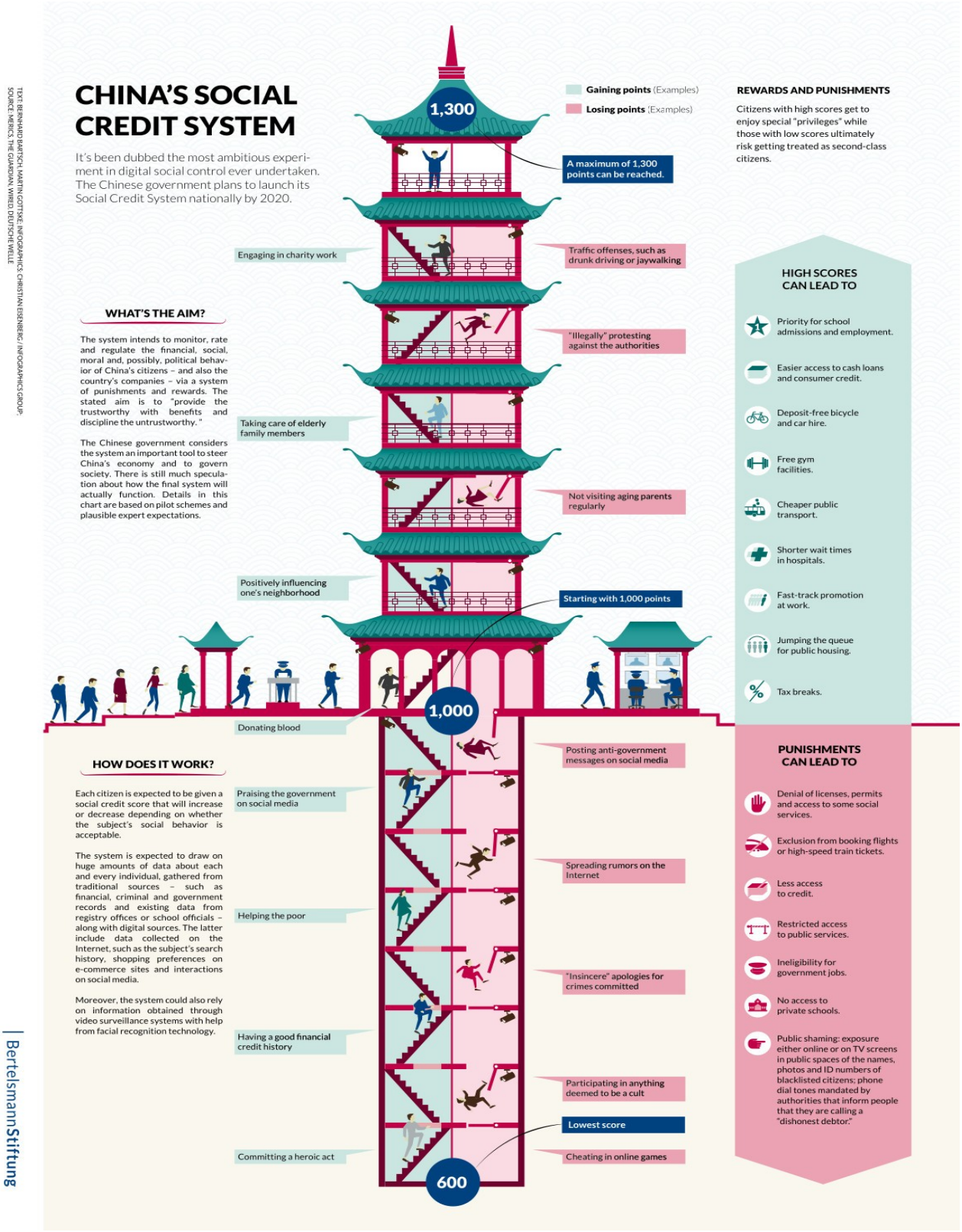
Kredit: AI4Diversity (https://twitter.com/AI4Diversity)

Príloha č. 2 – štruktúrované a neštruktúrované dáta



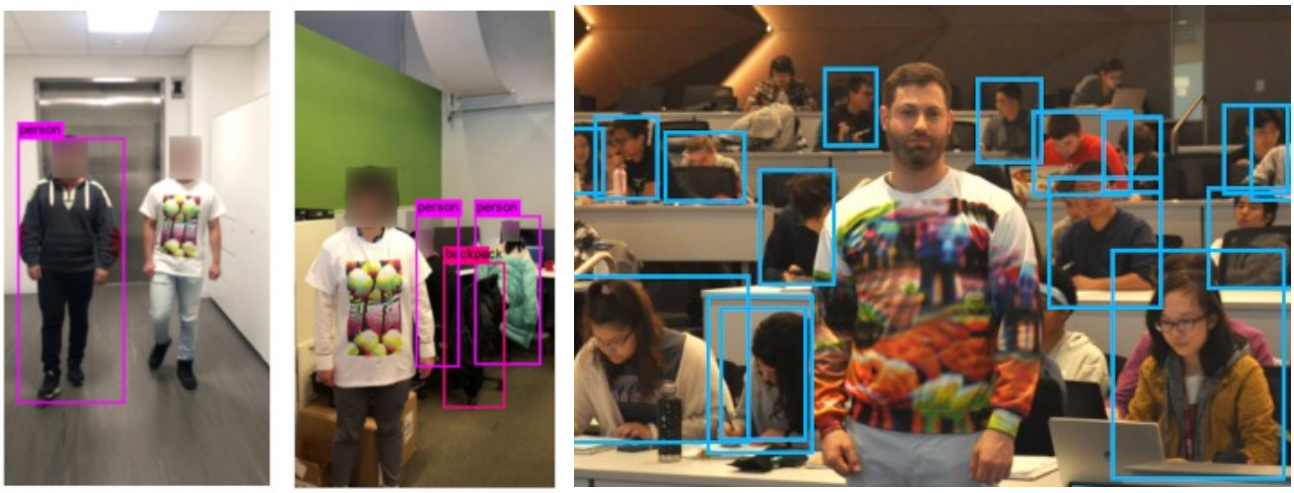
Kredit: AI4Diversity (https://twitter.com/AI4Diversity)

Príloha č. 3 – čínsky systém sociálneho kreditu

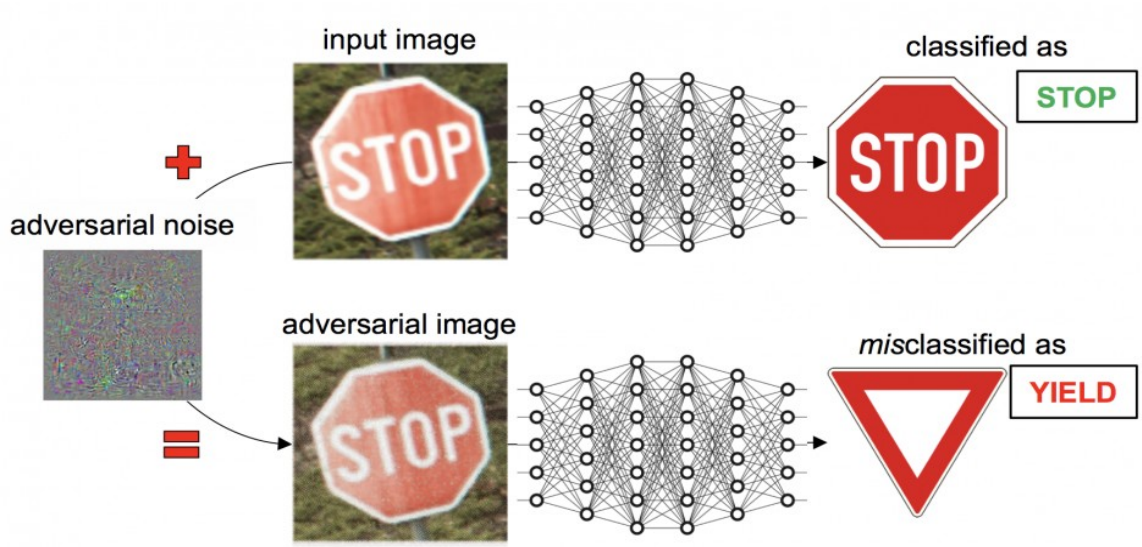


Kredit: Bertelsmann Stiftung (https://www.bertelsmann-stiftung.de/fileadmin/files/aam/Asia-Book_A_03_China_Social_Credit_System.pdf)

Príloha č. 4 – ďalšie príklady zraniteľností a klamania algoritmov AI



Kredit: XU, K. et al., WU, Z. et al.[668]



Kredit: AGARWAL, S.[669]

668 XU, K., ZHANG, G., LIU, S. et al. *Adversarial T-shirt! Evading Person Detectors in A Physical World*. ECVA, 2020. [on-line]. [cit. 20. júna 2022]. Dostupné na internete: <https://www.ecva.net/papers/eccv_2020/papers_ECCV/papers/123500647.pdf>
WU, Z, LIM, S., DAVIS, L., GOLDSTEIN, T. *Making an Invisibility Cloak: Real World Adversarial Attacks on Object Detectors*. ECVA, 2020. [on-line]. [cit. 20. júna 2022]. Dostupné na internete: <https://www.ecva.net/papers/eccv_2020/papers_ECCV/papers/123490001.pdf>

669 AGARWAL, S. *Machine learning Attack*. [on-line]. [cit. 20. júna 2022]. Dostupné na internete: <https://medium.com/mlearning-ai/machine-learning-attack-a92c5359b36d>

Zoznam použitej literatúry

ADIB-MOGHADDAM, A. *Artificial intelligence must not be allowed to replace the imperfection of human empathy*. [on-line]. [cit. 20. februára 2022].
Dostupné na internete: <https://theconversation.com/artificial-intelligence-must-not-be-allowed-to-replace-the-imperfection-of-human-empathy-151636>

AGARWAL, S. *Machine learning Attack*. [on-line]. [cit. 20. júna 2022].
Dostupné na internete: <https://medium.com/mlearning-ai/machine-learning-attack-a92c5359b36d>

AI Principles: Recommendations on the Ethical Use of Artificial Intelligence by the Department of Defense. [on-line]. [cit. 9. marca 2022].
Dostupné na internete: <https://admin.govexec.com/media/dib_ai_principles_-_supporting_document_-_embargoed_copy_(oct_2019).pdf>

AMODEI, D., OLAH, CH., STEINHARDT, J. et al. *Concrete Problems in AI Safety*. [on-line]. [cit. 7. júla 2022].
Dostupné na internete: <https://arxiv.org/abs/1606.06565>

ALFONSEC, M., CEBRIAN, M. et al. *Superintelligence Cannot be Contained: Lessons from Computability Theory.* [on-line]. [cit. 26. januára 2021].
Dostupné na internete: <https://jair.org/index.php/jair/article/view/12202>

Algoritmy strojového učenia I. [on-line]. [cit. 3. januára 2022].
Dostupné na internete: <*https://umelainteligencia.sk/algoritmy-strojoveho-ucenia/*>

Algoritmy strojového učenia II. [on-line]. [cit. 3. januára 2022].
Dostupné na internete: <https://umelainteligencia.sk/algoritmy-strojoveho-ucenia-ii-ucenie-bez-ucitela/>

Algoritmy strojového učenia III. [on-line]. [cit. 4. januára 2022].
Dostupné na internete: <https://umelainteligencia.sk/algoritmy-strojoveho-ucenia-iii-ucenie-formou-odmenovania/>

AMODEI, D., HERNANDEZ, D. *AI and Compute*. [on-line]. [cit. 19. januára 2022].
Dostupné na internete: <https://openai.com/blog/ai-and-compute/>

An Open Letter to the United Nations: Convention on Certain Conventional Weapons. [on-line]. [cit. 21. marca 2022].
Dostupné na internete: <https://futureoflife.org/2017/08/20/autonomous-weapons-open-letter-2017/>

ANDERSON, K., WAXMAN, M. C. *Law and Ethics for Autonomous Weapon Systems: Why a Ban Won't Work and How the Laws of War Can*. In: *Jean Perkins Task Force on National Security and Law Essay Series*. Stanford University: Hoover Institution Press, 2013.

ANDERSON, K., WAXMAN, M. C. *Law and Ethics for Robot Soldiers*. In: *Policy Review.* 2012, 176, 12.

Artificial Intelligence: Principles, laws, and frameworks. OneTrust DataGuidance Limited, 2022. ISSN 2398-9955.

ASIMOV, I. *Runaround*. Astounding Science Fiction, marec 1942.

Automated Vehicles for Safety. [on-line]. [cit. 7. júla 2022].
Dostupné na internete: <https://www.nhtsa.gov/technology-innovation/automated-vehicles-safety>

Autonómne vozidlá [on-line]. [cit. 6. augusta 2020].
Dostupné na internete: <http://akgunis.sk/autonomne-vozidla/>

Autonomous Weapons: An Open Letter from AI [Artificial Intelligence] & Robotics Researchers. [on-line]. Future of Life Institute, 2015. [cit. 8. marca 2022].
Dostupné na internete: <http://futureoflife.org/open-letter-autonomous-weapons/>

AUTOR, D., SALOMONS. A. *Is Automatization labor-displacing? Productivity growth, employment, and the labor share*. In: *Brookings Papers on Economic Activity*. [on-line]. Spring 2018, s. 1-63. [cit. 14. septembra 2022].
Dostupné na internete: <https://www.jstor.org/stable/26506212>

BECK, B. R. et al. *Predicting commercially available antiviral drugs that may act on the novel coronavirus (2019-nCoV), Wuhan, China through a drug-target interaction deep learning model*. In: *bioRxiv.* [online]. 2020, 2020.01.31.929547. [cit. 6. augusta 2020].
DOI: 10.1101/2020.01.31.929547
Dostupné na internete: <https://www.biorxiv.org/content/10.1101/2020.01.31.929547v1>

BENGIO, Y. *Machines Dream*. In: BEYER D. ed. *The Future of Machine Intelligence: Perspectives from Leading Practitioners*. Sebastopol, Calif.: O'Reilly Media, 2016.

BESSEN, J. *Artificial intelligence and jobs*. In: AGRAWAL, A., GANS, J., GOLDFARB, A. *The Economics of Artificial Intelligence: An Agenda*. [on-line]. 2019. [cit. 14. septembra 2022].
Dostupné na internete: <https://www.nber.org/papers/w24235>

BOGERT, E., SCHECTER, A., WATSON, R. T. *Humans rely more on algorithms than social influence as a task becomes more difficult*. In: *Sci Rep*. [on-line]. 2021, 11, 8028. [cit. 20. februára 2022].
DOI: 10.1038/s41598-021-87480-9

Dostupné na internete: <https://doi.org/10.1038/s41598-021-87480-9>

BOSTROM, N. *A history of transhumanist thought*. In: *Journal of Evolution and Technology.* [on-line]. 2005, roč. 14, vyd. 1. [cit. 8. januára 2022].
Dostupné na internete: <http://www.nickbostrom.com/papers/history.pdf>

BOSTROM, N. *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press, 2014. ISBN 978-0199678112.

CAMPBELL, J. *Dvojník.* Fantom Print, 2015. ISBN 978-80-7398-329-7.

CELLAN-JONES, R. *Stephen Hawking Warns Artificial Intelligence Could End Mankind*. In: *BBC News.* [on-line]. 2014, 2. 12. [cit. 5. augusta 2020].
Dostupné na internete: <https://www.bbc.com/news/technology-30290540>

CLAPPER, J. R. Jr. et al. *Unmanned Systems Roadmap: 2007-2032*. [on-line]. Washington, DC: Department of Defense [DOD], 2007. [cit. 5. marca 2022].
Dostupné na internete: <http://www.globalsecurity.org/intell/library/reports/2007/dod-unmanned-systems-roadmap_2007-2032.pdf>

CLARK, A. *Being There: Putting Brain, Body, and World Together Again*. Cambridge, Mass.: MIT Press, 1996.

Colonial Pipeline offline due to ransomware attack. [on-line]. [cit. 13. marca 2022].
Dostupné na internete: <https://www.fortiguard.com/outbreak-alert/darkside>

Compilation of open letters against autonomous weapons. [on-line]. [cit. 19. augusta 2020].
Dostupné na internete: <https://autonomousweapons.org/compilation-of-open-letters-against-autonomous-weapons/>

Conversations That Matter: The Crossroads of Science and Human Dignity Fall 2021. [on-line]. [cit. 28. marca 2022].
Dostupné na internete: <https://mcgrath.nd.edu/conferences/academic-pastoral/conversations-that-matter-the-crossroads-of-science-and-human-dignity/conversations-that-matter-the-crossroads-of-science-and-human-dignity-fall-2021/>

Cybercrime Is Now More Profitable Than The Drug Trade [on-line]. [cit. 10. marca 2022].
Dostupné na internete: <https://www.tripwire.com/state-of-security/regulatory-compliance/pci/cybercrime-is-now-more-profitable-than-the-drug-trade/>

Data Mining Techniques [on-line]. [cit. 7. decembra 2015].
Dostupné na internete: <http://documents.software.dell.com/statistics/textbook/data-mining-techniques>

Defense Science Board. *Task Force Report: The Role of Autonomy in DoD Systems*. Washington,

DC: Office of the Under Secretary of Defense for Acquisition, Technology and Logistics, 2012.

DOBBINS, J., COHEN, R. S., CHANDLER, N. et al. *Overextending and Unbalancing Russia: Assessing the Impact of Cost-Imposing Options.* [on-line]. Santa Monica, CA: RAND Corporation, 2019. [cit. 18. marca 2022].
Dostupné na internete: <https://www.rand.org/pubs/research_briefs/RB10014.html>

DOE/NNSA/LANL/SNL [on-line]. [cit. 3. marca 2022].
Dostupné na internete: <https://www.top500.org/site/50334/>

Dualism. [on-line]. [cit. 7. apríla 2022].
Dostupné na internete: <https://plato.stanford.edu/entries/dualism/>

Elements of AI. [on-line]. [cit. 30. marca 2022].
Dostupné na internete: <https://www.elementsofai.com/>

ELKUS, A. *How to be good: human values to artificial intelligence*. Slate, April 20, 2016.

Ethics guidelines for trustworthy AI. [on-line]. [cit. 28. marca 2022].
Dostupné na internete: <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>

ETZIONI, A., ETZIONI, O. *Keeping AI Legal*. In: *Vanderbilt Journal of Entertainment & Technology Law.* [on-line]. 2016, 19, no. 1, s. 133-146. [cit. 8. marca 2017].
Dostupné na internete: <http://www.jetlaw.org/wp-content/uploads/2016/12/Etzioni_Final.pdf>

ETZIONI, A., ETZIONI, O. *Pros and Cons of Autonomous Weapons Systems.* [on-line]. [cit. 5. marca 2022].
Dostupné na internete: <https://www.armyupress.army.mil/Journals/Military-Review/English-Edition-Archives/May-June-2017/Pros-and-Cons-of-Autonomous-Weapons-Systems/>

ETZIONI, O. *It's time to intelligently discuss artificial intelligence*. Backchannel, Dec 9, 2014.

FELDSTEIN, S. *The Global Expansion of AI Surveillance*. [on-line]. [cit. 28. februára 2022].
Dostupné na internete: <https://carnegieendowment.org/2019/09/17/global-expansion-of-ai-surveillance-pub-79847>

FERRIS, R. *Elon Musk thinks we will have to use AI this way to avoid a catastrophic future*. [on-line]. [cit. 24. januára 2022].
Dostupné na internete: <https://www.cnbc.com/2017/01/31/elon-musk-thinks-we-will-have-to-use-ai-this-way-to-avoid-a-catastrophic-future.html>

First Turing Test success marks milestone in computing history. [on-line]. [cit. 29. januára 2021].
Dostupné na internete: <https://phys.org/news/2014-06-turing-success-milestone-history.html>
FLYNN, J. *What Is Intelligence?: Beyond the Flynn Effect*. Cambridge University Press, 2009.

FORD, M. *Rise of the Robots: Technology and the Threat of a Jobless Future*. Basic Books, 2015.

CHADHA, S. *“Common Sense” is the Dark Matter of Artificial Intelligence*. [on-line]. [cit. 4. februára 2022].
Dostupné na internete: <https://hackernoon.com/the-dark-matter-of-ai-common-sense-is-not-so-common>

CHACE, C. *The Economic Singularity: Artificial Intelligence and the Death of Capitalism*. Three Cs, 2016.

GARAMONE, J. *9/11 Drove Change in Intelligence Community, NSA Chief Says*. [on-line]. [cit. 27. februára 2022].
Dostupné na internete: <https://www.defense.gov/News/News-Stories/Article/Article/945544/911-drove-change-in-intelligence-community-nsa-chief-says/>

Gartner Top Strategic Predictions For 2020 And Beyond. [on-line]. [cit. 23. marca 2022].
Dostupné na internete: <https://www.gartner.com/smarterwithgartner/gartner-top-strategic-predictions-for-2020-and-beyond>

GEIST, E., LOHN, A. J. *How Might Artificial Intelligence Affect the Risk of Nuclear War?* [on-line]. Santa Monica, CA: RAND Corporation, 2018. [cit. 21. marca 2022].
Dostupné na internete: <https://www.rand.org/pubs/perspectives/PE296.html>

GIBNEY, A. *Zero Days*. [filmový dokument]. [cit. 10. marca 2022].
Dostupné na internete: <http://www.zerodaysfilm.com/>

GIDDA, M. *Edward Snowden and the NSA files – timeline.* [on-line]. In: *The Gurdian*. 2013, 23. 6. [cit. 23. februára 2022].
Dostupné na internete: <http://www.guardian.co.uk/world/2013/jun/23/edward-snowden-nsa-files-timeline>

GOOD, I. J. *Speculations concerning the first untraintelligent machine*. In: *Advances in Computers*. Academic Press, vol. 6, 1965.

Google » Tensorflow: Vulnerability Statistics. [on-line]. [cit. 11. januára 2022].
Dostupné na internete: <https://www.cvedetails.com/product/53738/Google-Tensorflow.html?vendor_id=1224>

Google » Tensorflow: Security Vulnerabilities. [on-line]. [cit. 11. januára 2022].
Dostupné na internete: <https://www.cvedetails.com/vulnerability-list/vendor_id-1224/product_id-53738/Google-Tensorflow.html>

GREGOROVÁ, D. *Nedostatek spánku působí na mozek.* [on-line]. [cit. 20. januára 2022].
Dostupné na internete: <https://www.osel.cz/12127-nedostatek-spanku-pusobi-na-mozek.html>

GRIFFIN, A. *Facebook's artificial intelligence robots shut down after they start talking to each other in their own language* [on-line]. [cit. 6. augusta 2020].
Dostupné na internete:
<https://www.independent.co.uk/life-style/gadgets-and-tech/news/facebook-artificial-intelligence-ai-chatbot-new-language-research-openai-google-a7869706.html>

HAMILTON, I. A. *Amazon built an AI tool to hire people but had to shut it down because it was discriminating against women*. [on-line]. [cit. 6. augusta 2020].
Dostupné na internete: <https://www.businessinsider.com/amazon-built-ai-to-hire-people-discriminated-against-women-2018-10>

HAUGELAND J. *Artificial Intelligence: The Very Idea*. MIT Press, 1985.

HE, K., CHEN, X., XIE, S., LI, Y., DOLLÁR, P., GIRSHICK, R.: *Masked Autoencoders Are Scalable Vision Learners* [on-line]. Facebook AI Research (FAIR), 2021. [cit. 18. novembra 2021].
Dostupné na internete: <https://arxiv.org/abs/2111.06377>

HE, S. et al. *Learning to predict the cosmological structure formation.* [on-line]. In: *Proceedings of the National Academy of Sciences.* 2019, roč. 116, č. 28, s. 13825. [cit. 6. augusta 2020].
DOI: 10.1073/pnas.1821458116
Dostupné na internete: <https://www.pnas.org/content/116/28/13825>

HEATH N. *Google DeepMind founder Demis Hassabis: Three truths about AI* [on-line]. TechRepublic, September 24, 2018. [cit. 14. augusta 2022].
Dostupné na internete: <https://www.techrepublic.com/article/google-deepmind-founder-demis-hassabis-three-truths-about-ai/>

HIGGINS, A. *Stephen Hawking's final warning for humanity: AI is coming for us*. [on-line]. [cit. 13. marca 2022].
Dostupné na internete: <https://www.vox.com/future-perfect/2018/10/16/17978596/stephen-hawking-ai-climate-change-robots-future-universe-earth>

Hi Reddit, I'm Bill Gates and I'm back for my third AMA. Ask me anything. In: *Reddit*. [on-line]. 2015, 28. 1. [cit. 7. augusta 2020].
Dostupné na internete:
<https://www.reddit.com/r/IAmA/comments/2tzjp7/hi_reddit_im_bill_gates_and_im_back_for_my_third/?>

HOFSTADTER, D. R. *Analogy as the Core of Cognition*. [on-line]. Presidential Lecture, Stanford University, 2009. [cit. 7. apríla 2022].
Dostupné na internete: <https://www.youtube.com/watch?v=n8m7lFQ3njk>

HOFSTADTER, D. R. *Gödel, Escher, Bach: an Eternal Golden Braid*. New York: Basic Books, 1979. ISBN: 978-0-465-02656-2.

HOFSTADTER D. R. *Staring Emmy Straight in the Eye – and Doing My Best Not to Flinch*. In: DARTNELL T. *Creativity, Cognition, and Knowledge*. Westport, Conn.: Praeger, 2002.

HOFSTADTER, D. R, SANDER, E. *Surfaces and Essences*. New York: Basic Books, 2013.

How 9/11 Changed the Course of Personal Data Collection and Surveillance. [on-line]. [cit. 27. februára 2022].
Dostupné na internete: <https://www.startpage.com/privacy-please/startpage-articles/how-9-11-changed-the-course-of-personal-data-collection-and-surveillance>

HUMBER, M. *Technology and Workforce: Comparison between the Information Revolution and the Industrial Revolution* [on-line]. Berkeley: University of California, 2007. [cit. 20. augusta 2020].
Dostupné na internete: <http://infoscience.epfl.ch/record/146804/files/InformationSchool.pdf>

HUME, D. *A Treatise of Human Nature*. John Noon, 1738.

CHILD, A. *Origin of the Species*. In: *Apple TV.* [filmový dokument]. 2021. [cit. 25. apríla 2022].
Dostupné na internete:
<https://tv.apple.com/us/movie/origin-of-the-species/umc.cmc.1vsh8or9ojg824l4kn2kidkuw>

China wants to make its own rules for AI ethics. [on-line]. [cit. 21. marca 2022].
Dostupné na internete: <https://www.scmp.com/abacus/tech/article/3029194/china-wants-make-its-own-rules-ai-ethics>

CHOI, Q. CH. *Superintelligent AI May Be Impossible to Control; That's the Good News* [on-line]. [cit. 26. januára 2021].
Dostupné na internete: <https://spectrum.ieee.org/tech-talk/robotics/artificial-intelligence/super-artificialintelligence>

IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. [on-line]. [cit. 21. marca 2022].
Dostupné na internete: <http://standards.ieee.org/develop/indconn/ec/ead_v2.pdf>

IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. [on-line]. [cit. 31. marca 2022].
Dostupné na internete: <https://standards.ieee.org/industry-connections/ec/autonomous-systems/>

ITAPA: Umelá inteligencia v bežnom živote. [on-line]. [cit. 24. júna 2021].
Dostupné na internete: <https://www.itapa.sk/12826-sk/program/>

It's Getting Harder to Spot a Deep Fake Video. [on-line]. [cit. 21. marca 2022].
Dostupné na internete: <https://www.youtube.com/watch?v=gLoI9hAX9dw>

JÁN PAVOL II. *Redemptoris Missio*. Praha: Zvon, 1994.

JOBIN, A., IENCA, M., VAYENA, E. *The global landscape of AI ethics guidelines*. In: *Nat Mach Intell*. [on-line]. 2019, 1, s. 389–399. [cit. 28. marca 2022].
Dostupné na internete: <https://doi.org/10.1038/s42256-019-0088-2>

JOBIN, A., IENCA, M., VAYENA, E. *Artificial Intelligence: the global landscape of ethics guidelines*. [on-line]. [cit. 19. augusta 2020].
Dostupné na internete: <https://arxiv.org/pdf/1906.11668>

Katechizmus Katolíckej cirkvi. [on-line]. [cit. 10. apríla 2022].
Dostupné na internete: <https://katechizmus.sk/>

KELLY, K. *The Myth of a Superhuman AI*. In: *Wired*. [on-line]. 2017, 25. apríla. [cit. 7. septembra 2022].
Dostupné na internete: <https://www.wired.com/2017/04/the-myth-of-a-superhuman-ai/>

KHIMJI, I. *Cybercrime Is Now More Profitable Than The Drug Trade.* [on-line]. [cit. 14. februára 2022].
Dostupné na internete: <https://www.tripwire.com/state-of-security/regulatory-compliance/pci/cybercrime-is-now-more-profitable-than-the-drug-trade/>

KNIGHT, W. *The Dark Secret at the Heart of AI*. In: *Technology Review.* [on-line]. 2017, 11. 4.. [cit. 10. februára 2022].
Dostupné na internete: <https://www.technologyreview.com/2017/04/11/5113/the-dark-secret-at-the-heart-of-ai/>

KORAKOVOUNIS, D. *Spiking Neural Networks: where neuroscience meets artificial intelligence*. [on-line]. [cit. 27. februára 2022].
Dostupné na internete: <https://theaisummer.com/spiking-neural-networks/>

KOUSHIK, J. *Understanding Convolutional Neural Networks*. [on-line]. [cit. 31. januára 2022].
Dostupné na internete: <https://arxiv.org/abs/1605.09081>

KOZA, J. *Genetic Programming: On the Programming of Computers by Means of Natural Selection*. MIT Press, 1992.

KRAFT, A. *Microsoft shuts down AI chatbot after it turned into a Nazi*. [on-line]. [cit. 6. augusta 2020].
Dostupné na internete: <https://www.cbsnews.com/news/microsoft-shuts-down-ai-chatbot-after-it-turned-into-racist-nazi/>

LAM, L., CHEN, S. *US spies on Chinese mobile phone companies, steals SMS data: Edward Snowden* In: *South China Morning Post.* [on-line]. 2013, 23. 3. [cit. 23. februára 2022].

Dostupné na internete: <https://www.scmp.com/news/china/article/1266821/us-hacks-chinese-mobile-phone-companies-steals-sms-data-edward-snowden>

LAYTON, P. *Fighting Artificial Intelligence Battle: Operational Concept for Future AI-Enabled Wars*. [on-line]. Australian Defence College, 2021. [cit. 10. októbra 2022].
Dostupné na internete: <https://tasdcrc.com.au/wp-content/uploads/2021/02/JSPS_4.pdf>

LECUN, Y., MISRA, I. *Self-supervised learning: The dark matter of intelligence* [on-line]. [cit. 4. februára 2022].
Dostupné na internete: <https://ai.facebook.com/blog/self-supervised-learning-the-dark-matter-of-intelligence>

LEE, I. *Equalism: Paradise Regained*. In: LEE N. (ed.). *The Transhumanist Handbook*. Springer, 2019, s. 849 – 863.

LEHMAN, J., CLUNE, J., RISI, S.: *An Anarchy of Methods: Current Trends in How Intelligence Is Abstracted in AI*. In: IEEE Intelligent Systems. 2014, 29, č. 6.

Louis Pasteur citáty. [on-line]. [cit. 7. júla 2022].
Dostupné na internete: <https://citaty-slavnych.sk/autori/louis-pasteur/>

MAASS, W. *Networks of spiking neurons: The third generation of neural network models*. [on-line]. [cit. 28. februára 2022].
Dostupné na internete:
<https://www.sciencedirect.com/science/article/abs/pii/S0893608097000117>

MARCUS, G. *Deep Learning: A Critical Appraisal*. [on-line]. [cit. 7. apríla 2022].
Dostupné na internete: <https://arxiv.org/abs/1801.00631>

MARCHANT, G. E. et al. *International Governance of Autonomous Military Robots*. In: Columbia Science and Technology Law Review. [on-line]. 2011, 12. 6., s. 272–276. [cit. 27. marca 2017].
Dostupné na internete: <http://stlr.org/download/volumes/volume12/marchant.pdf>

MCCARTHY, J. et al. *Proposal for the Dartmouth Summer Research Project in Artificial Intelligence*. In: *AI Magazine*, [on-line]. 1955, 27(4). [cit. 3. februára 2021].
Dostupné na internete: <https://doi.org/10.1609/aimag.v27i4.1904>

MCFARLAND, M. *Elon Musk: „With Artificial Intelligence, We Are Summoning the Demon“*. In: *Washington Post.* 2014, 24. 10..

Measuring trends in Artificial Intelligence – 2021 AI Index Report. [on-line]. [cit. 28. marca 2022].
Dostupné na internete: <https://aiindex.stanford.edu/ai-index-report-2021/>

MERCER, C., TROTHEN, T. J. *Religion and Transhumanism*: The Unknown Future of Human Enhancement. Praeger, 2014.

METZ, C. *A New Way for Machines to See, Taking Shape in Toronto*. New York Times, Nov. 28, 2017.

MIHULKA, S. *První varování před koronavirem Wuhan poslala umělá inteligence BlueDot*. [on-line]. [cit. 6. augusta 2020].
Dostupné na internete: <https://www.osel.cz/11002-prvni-varovani-pred-koronavirem-wuhan-poslala-umela-inteligence-bluedot.html>

MIHULKA, S. *Letecká bojová umělá inteligence si natřela na chleba taktické experty*. [on-line]. [cit. 5. marca 2022].
Dostupné na internete: <https://www.osel.cz/8903-letecka-bojova-umela-inteligence-si-natrela-na-chleba-takticke-experty.html>

MINSKY, M. L. *Computation: Finite and Infinite Machines*. Upper Saddle River, N.J.: Prentice Hall, 1967.

MINSKY, M. L. *The Emotion Machine: Commonsense Thinking, Artificial Intelligence, and the Future of the Human Mind*. New York: Simon & Schuster, 2006.

MINSKY, M. L., PAPERT, S. L. *Perceptrons: An Introduction to Computational Geometry*. Cambridge, Mass: MIT Press, 1969.

MINSKY, M. *Society of Mind*. New York: Simon & Schuster, 1986.

MITCHELL, M. *An Introduction to Genetic Algorithms*. Cambridge, Mas: MIT Press, 1996.

MITCHELL, M. *Artificial Intelligence*. Farrar, Straus and Giroux, 2019, ISBN: 978-0-374-71523-6.

MITCHELL, M. *Conceptual Abstraction and Analogy in Artificial Intelligence*. In: *ALIFE 2020: The 2020 Conference on Artificial Life*. [on-line]. 2020. [cit. 5. apríla 2022].
Dostupné na internete: <https://doi.org/10.1162/isal_a_00354>

MOEWES, C., NÜRNBERGER, A. *Computational Intelligence in Intelligent Data Analysis*. New York: Springer, 2013.

Moral Machine. [on-line]. [cit. 7. júla 2022].
Dostupné na internete: <https://www.moralmachine.net/>

MORAVEC, H. *Mind Children: The Future of Robot and Human Intelligence*. Cambridge, Mass.: Hardvard University Press, 1988.

MORACEC, H. *Robot: Mere Machine to Transcendent Mind*. Oxford University Press, 2000.

MUEHLHAUSER, L. *Facing the Intelligence Explosion*. MIRI, 2013.

MÜLLER, V. C., BOSTROM, N. *Future Progress in Artificial Intelligence: A Survey of Expert Opinion*. In: *Fundamental Issues of Artificial Intelligence. Cham*, Switzerland: Springer

International, 2016, s. 555-572.

Nariadenie Európskeho parlamentu a Rady (EÚ) 2016/679 z 27. apríla 2016 o ochrane fyzických osôb pri spracúvaní osobných údajov a o voľnom pohybe takýchto údajov, ktorým sa zrušuje smernica 95/46/ES (všeobecné nariadenie o ochrane údajov). [on-line]. [cit. 19. februára 2022]. Dostupné na internete: <https://eur-lex.europa.eu/legal-content/SK/TXT/?uri=celex%3A32016R0679>

Nariadenie Európskeho parlamentu a Rady, ktorým sa stanovujú harmonizované pravidlá v oblasti umelej inteligencie (Akt o umelej inteligencii) a menia niektoré legislatívne akty únie. [on-line]. [cit. 24. marca 2022].
Dostupné na internete: <https://eur-lex.europa.eu/legal-content/SK/TXT/?uri=CELEX:52021PC0206>

New Deep Learning Method Adds 301 Planets to Kepler's Total Count. [on-line]. [cit. 28. novembra 2021].
Dostupné na internete: <https://www.jpl.nasa.gov/news/new-deep-learning-method-adds-301-planets-to-keplers-total-count>

NG, A. *Deep Learning in Practice: Speech Recognition and Beyond*. In: *EmTech Digital.* [on-line]. 2016, 23. 5. [cit. 3. februára 2022].
Dostupné na internete: <https://events.technologyreview.com/video/watch/andrew-ng-deep-learning/>

NGUYEN, A., YOSINSKI, J., CLUNE, J. *Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images*. [on-line]. CVPR, 2015. [cit. 12. februára 2022]. Dostupné na internete: <https:// cv-foundation.org/openaccess/content_cvpr_2015/papers/Nguyen_Deep_Neural_Networks_2015_CVPR_paper.pdf>

NIELSEN, M. *Neural Networks and Deep Learning*. [on-line]. [cit. 19. januára 2022].
Dostupné na internete: <http://neuralnetworksanddeeplearning.com/>

NILSSON, N. J., MCCARTHY, J. *A Biographical Memoir*. Washington D.C.: National Academy of Sciences, 2012.

OMOHUNDRO, S. *The basic AI drives*. In: *Artificial General Intelligence 2008: Proceedings of the First AGI Conference*. IOS Press, 2008.

One Hundred Year Study on Artificial Intelligence. [on-line]. AI100, 2016. [cit. 14. novembra 2021].
Dostupné na internete: <https://ai100.stanford.edu/2016-report>

One Hundred Year Study on Artificial Intelligence. [on-line]. AI100, 2021. [cit. 14. novembra 2021].
Dostupné na internete: <https://ai100.stanford.edu/2021-report>

Open Letter on Research Priorities for Robust and Beneficial Artificial Intelligence. [on-line]. [cit. 21. marca 2022].
Dostupné na internete: <https://futureoflife.org/2015/10/27/ai-open-letter/>

ORLOWSKI, J. *The Social Dilemma*. Netflix, 2020. [filmový dokument]. [cit. 7. decembra 2021].
Dostupné na internete: <https://www.netflix.com/sk/title/81254224>

От робототехники до беспилотников: Шойгу рассказал о новинках ОПК на форуме «Армия-2020». [on-line]. [cit. 1. marca 2022].
Dostupné na internete: <https://youtu.be/U8V4OZyNSac?t=152>

PAROULKOVÁ, V. *Lidská práva pro roboty? Evropská unie chystá právní statut tzv. umělé osoby* [on-line]. [cit. 6. augusta 2020].
Dostupné na internete: <https://plus.rozhlas.cz/lidska-prava-pro-roboty-evropska-unie-chysta-pravni-statut-tzv-umele-osoby-6598059>

PATTYNOVÁ, J. *Výzvy a právní aspekty umělé inteligence*. In: *Umělá inteligence 2021*. Praha: TUESDAY Business Network, 2021.
Dostupné na internete: <https://www.tuesday.cz/akce/umela-inteligence-1/zaznam-akce/>

PEARL, J., MACKENZIE, D. *The Book of Why: THe New Science of Cause and Effect*. New York: Basic Books, 2018.

PFEIFFER, M., PFEIL, T. *Deep Learning With Spiking Neurons: Opportunities and Challenges*. [on-line]. [cit. 27. februára 2022].
Dostupné na internete: <https://www.frontiersin.org/articles/10.3389/fnins.2018.00774/full>

POITRAS, L., ROSENBACH, M., SCHMID, F., STARK, H. *NSA horcht EU-Vertretungen mit Wanzen aus* In: *Der Spiegel.* [on-line]. 2013, 29. 6. [cit. 23. februára 2022].
Dostupné na internete: <http://www.spiegel.de/netzwelt/netzpolitik/nsa-hat-wanzen-in-eu-gebaeuden-installiert-a-908515.html>

POITRAS, L., ROSENBACH, M., STARK, H. *NSA überwacht 500 Millionen Verbindungen in Deutschland* In: *Der Spiegel*. [on-line]. 2013, 30. 6. [cit. 23. februára 2022].
Dostupné na internete: <http://www.spiegel.de/netzwelt/netzpolitik/nsa-ueberwacht-500-millionen-verbindungen-in-deutschland-a-908517.html>

Prečo Industry 4.0 [on-line]. [cit. 20. augusta 2020].
Dostupné na internete: <http://industry4.sk/o-industry-4-0/co-je-industry-4-0/>

PRESS, G. *12 Observations About Artificial Intelligence From The O'Reilly AI Conference*. In: *Forbes*. [on-line]. 2016, 31. 10. [cit. 7. augusta 2020].
Dostupné na internete: <https://www.forbes.com/sites/gilpress/2016/10/31/12-observations-about-

artificial-intelligence-from-the-oreilly-ai-conference/>

Profile: Edward Snowden In: *BBC.* [on-line]. 2013, 24. 6. [cit. 23. februára 2022].
Dostupné na internete: <http://www.bbc.co.uk/news/world-us-canada-22837100>

REHÁK, M. *Útoky na systémy umělé inteligence a jejich obrana*. In: *Umělá inteligence 2021*. Praha: TUESDAY Business Network, 2021.
Dostupné na internete: <https://www.tuesday.cz/akce/umela-inteligence-1/zaznam-akce/>

RHODES, R. *The Making of the Atomic Bomb*. Simon & Schuster, 1987.

Robot Sophia speaks at Saudi Arabia's Future Investment Initiative [on-line]. [cit. 6. augusta 2020].
Dostupné na internete: <https://youtu.be/dMrX08PxUNY>

Rome Call for AI Ethics. [on-line]. [cit. 19. augusta 2020].
Dostupné na internete: <http://www.academyforlife.va/content/pav/en/events/intelligenza-artificiale.html>

Rome Call for AI Ethics (document). [on-line]. [cit. 28. marca 2022].
Dostupné na internete:
<https://www.romecall.org/wp-content/uploads/2022/03/RomeCall_Paper_web.pdf>

ROMPORTL, J. *Umělá inteligence, Life 3.0, superinteligence a život ve vesmíru*. [prednáška]. YouTube, 2019. [cit. 7. júla 2022].
Dostupné na internete: <https://www.youtube.com/watch?v=PaLrrczJ5VI>

ROSENBLATT, F. *The Perceptron: A Probabilistic Model of Information Storage and Organization in the Brain*. In: *Psychological Review*. 1958, 65, č. 6.

ROTA, G.C. *Indiscrete Thoughts*. Boston: Berkhäuser, 1997.

RUMELHART, D. E., MCCLELLAND, J. L. *Parallel Distributed Processing.* Vol 1/2. Bradford Book, 1986.

RUSSELL, S. *Human Compatible*. Penguin Books, 2020, ISBN: 978-0-241-33524-6.

SAMPLE, I. *Thousands of leading AI researchers sign pledge against killer robots*. [on-line]. [cit. 21. marca 2022].
Dostupné na internete: <https://www.theguardian.com/science/2018/jul/18/thousands-of-scientists-pledge-not-to-help-build-killer-ai-robots>

SAMUEL, A. L. *Some Studies in Machine Learning Using the Game of Checkers*. In: *IBM Journal of Research and Development.* 1959, č. 3.

SANDBERG, A. *Whole brain emulation: A roadmap*. Future of Humanity Institute. Oxford University, 2008.

Scientists' Call to Ban Autonomous Lethal Robots. ICRAC, October 2013. [on-line]. [cit. 8. marca 2022].
Dostupné na internete: <http://www.icrac.net/>

SEARLE, R. J. *Minds, Brains, and Programs* In: *The Behavioral and Brain Sciences*. [on-line]. Cambridge University Press, 1980, zv. 3. [cit. 30. januára 2021].
Dostupné na internete: <https://web.archive.org/web/20071210043312/http://members.aol.com/NeoNoetics/MindsBrainsPrograms.html>

SHARKEY, N. *Saying 'No!' to Lethal Autonomous Targeting*. In: *Journal of Military Ethics.* 2010, 9, č. 4, s. 369–383.

SHENKER I. *Brainy robots in our future, experts think*. Detroit Free Press, September 30, 1977.

Scholastic Method [on-line]. [cit. 6. novembra 2021].
Dostupné na internete: <https://www.encyclopedia.com/religion/encyclopedias-almanacs-transcripts-and-maps/scholastic-method>

SIEGLER, M. G. *Eric Schmidt: Every 2 Days We Create As Much Information As We Did Up To 2003* [on-line]. [cit. 12. decembra 2021].
Dostupné na internete: <http://techcrunch.com/2010/08/04/schmidt-data/>

SIMON, H. A. *Artificial Intelligence: An Empirical Science*. In: *Artificial Intelligence.* 1955, 77, č. 2.

SIMON, H. A. *The Shape of Automation for Men and Management*. New York: Harper & Row, 1965.

SMITH, R. *Why No Science?* [on-line]. [cit. 6. novembra 2021].
Dostupné na internete: <https://www.thecatholicthing.org/2021/11/02/why-no-science/>

SOFGE, E. *Bill Gates fears AI, but AI researchers know better*. Popular Science, Jan 30, 2015.

SPARROW, R. *Killer Robots*. In: *Journal of Applied Philosophy.* 2007, 24, č. 1, s. 62–77.

STRAHOVNIK, V. *Virtues and transhumanist human enhancement*. In: PETROUŠEK, R., ŽALEC, B. *Transhumanism as a Chalange for Ethics and Religion*. 2021, s. 37 – 44.

Stratégia kybernetickej obrany Slovenskej republiky. [on-line]. Ministerstvo obrany SR, Vojenské spravodajstvo. [cit. 17. marca 2022].
Dostupné na internete: <https://www.slov-lex.sk/legislativne-procesy/-/SK/dokumenty/LP-2022-128>

SUMMER, T. *The first AI universe sim is fast and accurate—and its creators don't know how it works*. [on-line]. [cit. 6. augusta 2020].
Dostupné na internete: <https://phys.org/pdf480780725.pdf>

SZEGEDY, CH. et al. *Intriguing Properties of Neural Networks*. In: *Proceedings of the International Conference on Learning Representations*. 2014.

SZONDY, D. *General Atomics' Gambit autonomous combat drone takes the initiative*. [on-line]. [cit. 7. marca 2022].
Dostupné na internete: <https://newatlas.com/military/general-atomics-gambit-combat-drone-air-dominance/>

SZONDY, D. *Skyborg combat drone's "brain" flies for the first time*. [on-line]. [cit. 7. marca 2022].
Dostupné na internete: <https://newatlas.com/military/skyborg-combat-drones-brain-first-time/>

ŠANTAVÝ, P. *Analýza virtuálneho sveta v kontexte náboženskej formácie, pôsobenia a života Cirkvi* [licenciátska práca]. [on-line]. Bratislava: RKCMBF UK, 2017. [cit. 19. augusta 2020 – 7. decembra 2021].
Dostupné na internete:
<https://peter.santavy.cloud/data/uploads/docs/analyza_virtualneho_sveta_v_kontexte_nabozenskej_formacie_posobenia_a_zivota_cirkvi.pdf>

ŠANTAVÝ, P. *Etické výzvy a morálne aspekty súčasných systémov umelej inteligencie* [dizertačná práca]. [on-line]. Bratislava: RKCMBF UK, 2022.
Dostupné na internete:
<https://peter.santavy.cloud/data/uploads/docs/eticke_vyzvy_a_moralne_aspekty_sucasnych_systemov_umelej_inteligencie.pdf>

ŠANTAVÝ, P. Dejiny spásy: zmluvy Starého zákona a Nová zmluva [diplomová práca]. [on-line]. Bratislava: RKCMBF UK, 2000. [cit. 6. novembra 2021].
Dostupné na internete:
<https://peter.santavy.cloud/data/uploads/docs/dejiny_spasy-zmluvy_sz_a_nz.pdf>

ŠANTAVÝ, P. *Niektoré výzvy informačnej spoločnosti v oblasti morálnej teológie*. In *Doctorandum dies 2018: Varia historia et moralia*. RKCMBF UK, Bratislava 2018.

ŠTARHA, Š., GAŠPAROVIČ, R. *AI z pohľadu práva*. [on-line]. [cit. 29. marca 2022].
Dostupné na internete: <https://www.epravo.sk/top/clanky/ai-z-pohladu-prava-4483.html>

TANZ, J. *Soon We Won't Program Computers. We'll Train Them Like Dogs*. Wired, May 17, 2016.

The AI arms race. [on-line]. [cit. 21. marca 2022].
Dostupné na internete: <https://www.ft.com/content/21eb5996-89a3-11e8-bf9e-8771d5404543>

The AI Powered State. [on-line]. [cit. 26. marca 2022].
Dostupné na internete: <https://www.nesta.org.uk/feature/ai-powered-state/>

The „good“ algorithm. [on-line]. [cit. 28. marca 2022].

Dostupné na internete: <https://www.academyforlife.va/content/dam/pav/documenti%20pdf/2020/Assemblea/Atti_Assemblea_e_28febbraio/Atti%20completi_PAV_2020_.pdf>

The Myth Of AI: A Conversation With Jaron Lanier. [on-line]. [cit. 7. júla 2022].
Dostupné na internete: <https://www.edge.org/conversation/jaron_lanier-the-myth-of-ai>

The Nonhuman Rights Project: Frequently Asked Questions. [on-line]. [cit. 10. apríla 2022].
Dostupné na internete: <https://www.nonhumanrights.org/frequently-asked-questions/>

The People's Liberation Army Strategic Support Force. [on-line]. [cit. 13. marca 2022].
Dostupné na internete: <https://jamestown.org/program/the-peoples-liberation-army-strategic-support-force-update-2019/>

The Price of Privacy: Re-Evaluating the NSA [on-line]. [cit. 23. februára 2022].
Dostupné na internete: <https://www.youtube.com/watch?v=kV2HDM86XgI&t=18m>

The Turing Digital Archive [on-line]. [cit. 30. januára 2021].
Dostupné na internete: <http://www.turingarchive.org/>

THURZO, V. *The Influence of Existentialism and Subjectivism on the Concept of the Human Person*. In: *Spiritual and Social Experience in the Context of Modernism and Postmodernism*. Morrisville, 2021.

Top Strategic Predictions for 2020 and Beyond: Technology Changes the Human Condition. [on-line]. [cit. 24. marca 2022].
Dostupné na internete: <https://www.gartner.com/document/3970846>

THOMPSON, N. C., GREENEWALD, K., LEE, K., MANSO, G. F. *Deep learning computational cost*. [on-line]. [cit. 27. februára 2022].
Dostupné na internete: <https://spectrum.ieee.org/deep-learning-computational-cost>

THORTON, R. *The age of machinery*. Primitive Expounder IV, 1847, s. 281.

THURNHER, J. S. *Legal Implications of Fully Autonomous Targeting*. In: *Joint Force Quarterly*. [on-line]. 2012, 67, 4, s. 83. [cit. 7. marca 2022].
Dostupné na internete: <http://ndupress.ndu.edu/Portals/68/Documents/jfq/jfq-67/JFQ-67_77-84_Thurnher.pdf>

Transhumanist Declaration. [on-line]. [cit. 26. januára 2022].
Dostupné na internete: <https://hpluspedia.org/wiki/Transhumanist_Declaration>

TROTT, D. *The hard things are easy, but the easy things are hard.* [on-line]. [cit. 8. januára 2022].
Dostupné na internete: <https://www.campaignlive.com/article/hard-things-easy-easy-things-hard/1498154>

Trustworthy AI is human-centered. [on-line]. [cit. 20. februára 2022].
Dostupné na internete: <https://www.ibm.com/watson/trustworthy-ai>

TUCKER, P. *SecDef: China Is Exporting Killer Robots to the Mideast*. [on-line]. [cit. 6. decembra 2021].
Dostupné na internete: <https://www.defenseone.com/technology/2019/11/secdef-china-exporting-killer-robots-mideast/161100/>

TUCKER, P. *The Pentagon's AI Ethics Draft Is Actually Pretty Good*. [on-line]. [cit. 9. marca 2022].
Dostupné na internete: <https://www.defenseone.com/technology/2019/10/pentagons-ai-ethics-draft-actually-pretty-good/161005/>

TURING, A. *Computing machinery and intelligence*. Mind 59, 1950, s. 433-460.

TURING, A. *Intelligent machinery, a heretical theory*. The 51 Society, Manchester, 1951.

Understanding China's AI Strategy. [on-line]. [cit. 21. marca 2022].
Dostupné na internete: <https://www.cnas.org/publications/reports/understanding-chinas-ai-strategy>

URBINA, F., LENTZOS, F., INVERNIZZI, C. et al. *Dual use of artificial-intelligence-powered drug discovery*. In: *Nat Mach Intell*. [on-line]. 2022, 4, s. 189–191. [cit. 22. marca 2022].
Dostupné na internete: <https://doi.org/10.1038/s42256-022-00465-9>

Uznesenie Európskeho parlamentu zo dňa 16. 2. 2017 obsahujúce odporúčania pre Komisiu k normám občianskeho práva v oblasti robotiky (2015/2103(INL)). [on-line]. [cit. 29. marca 2022].
Dostupné na internete: <https://eur-lex.europa.eu/legal-content/SK/TXT/PDF/?uri=CELEX:52017IP0051&from=EN>

Vatican Hackathon – harnessing youth, technology to serve common good. [on-line]. [cit. 28. marca 2022].
Dostupné na internete: <https://www.vaticannews.va/en/vatican-city/news/2018-03/vatican-hackathon--.html>

VIGNARD, K. *Manifestos and open letters: Back to the future?* [on-line]. [cit. 21. marca 2022].
Dostupné na internete: <https://thebulletin.org/2018/04/manifestos-and-open-letters-back-to-the-future/>

Viktor Frankl citáty. [on-line]. [cit. 25. apríla 2022].
Dostupné na internete: <https://citaty-slavnych.sk/autori/viktor-frankl/>

VIVODA, M. *Transhumanizmus a Katolícka Cirkev*. In: *Nové Horizonty*. 2021, roč. 15, č. 3, s. 121-128. ISSN: 1337-6535, EAN 977133765400605.

Všeobecná deklarácia ľudských práv. [on-line]. New York, 10.12.1948.

Dostupné na internete: <https://www.ustavnysud.sk/documents/10182/992240/DE01_48.pdf/>

‘Wake up’: Ex-Air Force software officer warns China is winning AI battle. [on-line]. [cit. 21. marca 2022].
Dostupné na internete: <https://www.washingtontimes.com/news/2021/oct/11/nicolas-chaillan-warns-china-winning-ai-battle/>

Why Advanced Ransomware Is Cybercrime's Most Profitable Business Model [on-line]. [cit. 10. marca 2022].
Dostupné na internete: <https://blog.knowbe4.com/why-advanced-ransomware-is-cybercrimes-most-profitable-business-model>

WikiLeaks Releases Trove of Alleged C.I.A. Hacking Documents. [on-line]. [cit. 27. februára 2022].
Dostupné na internete: <https://www.nytimes.com/2017/03/07/world/europe/wikileaks-cia-hacking.html>

WIENER, N. *God and Golem, Inc.: A Comment on Certain Points Where Cybernetics Impinges on Religion*. MIT Press, 1964.

WIENER, N. *The Human Use of Human Beings*. Riverside Press, 1950.

WOODS, E. T. Jr. *Ako Katolícka cirkev budovala západnú civilizáciu*. Bratislava: Redemptoristi - Slovo medzi nami, 2010.

WU, Z, LIM, S., DAVIS, L., GOLDSTEIN, T. *Making an Invisibility Cloak: Real World Adversarial Attacks on Object Detectors*. ECVA, 2020. [on-line]. [cit. 20. júna 2022].
Dostupné na internete:
<https://www.ecva.net/papers/eccv_2020/papers_ECCV/papers/123490001.pdf>

XU, K., ZHANG, G., LIU, S. et al. *Adversarial T-shirt! Evading Person Detectors in A Physical World*. ECVA, 2020. [on-line]. [cit. 20. júna 2022].
Dostupné na internete:
<https://www.ecva.net/papers/eccv_2020/papers_ECCV/papers/123500647.pdf>

YUDKOWSKY, E. *Intelligence explosion microeconomics*. MIRI, 2013.

ZUBOFF, SH. *The real reason why Facebook and Google won’t change*. [on-line]. [cit. 21. marca 2022].
Dostupné na internete: <https://www.theguardian.com/science/2018/jul/18/thousands-of-scientists-pledge-not-to-help-build-killer-ai-robots>

Twitter:

BEYER, L. *The return of patch-based self-supervision...* [on-line]. [cit. 18. novembra 2021]. Dostupné na internete: <https://twitter.com/giffmana/status/1459092079020285976>

KARPATHY, A. *Computer vision research feels a bit stagnating in a local minimum of 2D texture recognition on ImageNet...* [on-line]. [cit. 10. februára 2022]. Dostupné na internete: <https://twitter.com/karpathy/status/1491452689825165314>

Wikipedia.org:

5-HT$_{2A}$ receptor. [on-line]. [cit. 20. januára 2022]. Dostupné na internete: <https://en.wikipedia.org/wiki/5-HT2A_receptor>

Action potential [on-line]. [cit. 27. februára 2022]. Dostupné na internete: <https://en.wikipedia.org/wiki/Action_potential>

Alan Turing. [on-line]. [cit. 30. januára 2021]. Dostupné na internete: <https://en.wikipedia.org/wiki/Alan_Turing#Pattern_formation_and_mathematical_biology>

AlphaGo. [on-line]. [cit. 30. júla 2020]. Dostupné na internete: <https://en.wikipedia.org/wiki/AlphaGo>

Carnivore [on-line]. [cit. 23. februára 2022]. Dostupné na internete: <https://en.wikipedia.org/wiki/Carnivore_(software)>

Common Vulnerabilities and Exposures. [on-line]. [cit. 15. februára 2022]. Dostupné na internete: <hhttps://en.wikipedia.org/wiki/Common_Vulnerabilities_and_Exposures>

Darknet [on-line]. [cit. 23. februára 2022]. Dostupné na internete: <https://en.wikipedia.org/wiki/Darknet>

Dartmouth workshop. [on-line]. [cit. 1. februára 2021]. Dostupné na internete: <https://en.wikipedia.org/wiki/Dartmouth_workshop>

Douglas Hofstadter. [on-line]. [cit. 30. júla 2020]. Dostupné na internete: <https://en.wikipedia.org/wiki/Douglas_Hofstadter>

Deep Blue versus Garry Kasparov. [on-line]. [cit. 30. júla 2020]. Dostupné na internete: <https://en.wikipedia.org/wiki/Deep_Blue_versus_Garry_Kasparov>

Echelon [on-line]. [cit. 23. februára 2022].

Dostupné na internete: <https://en.wikipedia.org/wiki/ECHELON>

Equation Group. [on-line]. [cit. 10. marca 2022].
Dostupné na internete: <https://en.wikipedia.org/wiki/Equation_Group>

Ex Machina (film). [on-line]. [cit. 30. septembra 2022].
Dostupné na internete: <https://en.wikipedia.org/wiki/Ex_Machina_(film)>

Fourth-generation warfare. [on-line]. [cit. 10. marca 2022].
Dostupné na internete: <https://en.wikipedia.org/wiki/Fourth-generation_warfare>

Global surveillance [on-line]. [cit. 23. februára 2022].
Dostupné na internete: <https://en.wikipedia.org/wiki/Global_surveillance>

HermeticWiper: New data-wiping malware hits Ukraine. [on-line]. [cit. 10. marca 2022].
Dostupné na internete: <https://www.welivesecurity.com/2022/02/24/hermeticwiper-new-data-wiping-malware-hits-ukraine/>

IsaacWiper and HermeticWizard: New wiper and worm targeting Ukraine. [on-line]. [cit. 10. marca 2022].
Dostupné na internete: <https://www.welivesecurity.com/2022/03/01/isaacwiper-hermeticwizard-wiper-worm-targeting-ukraine/>

Kybernetika. [on-line]. [cit. 3. februára 2021].
Dostupné na internete: <https://sk.wikipedia.org/wiki/Kybernetika>

Lethal autonomous weapon. [on-line]. [cit. 5. marca 2022].
Dostupné na internete: <https://en.wikipedia.org/wiki/Lethal_autonomous_weapon>

Missionaries and cannibals problem. [on-line]. [cit. 18. novembra 2021].
Dostupné na internete:
<https://en.wikipedia.org/wiki/Missionaries_and_cannibals_problem>

Neuralink. [on-line]. [cit. 5. augusta 2020].
Dostupné na internete: <https://en.wikipedia.org/wiki/Neuralink>

Occam's Razor. [on-line]. [cit. 6. novembra 2021].
Dostupné na internete: <https://en.wikipedia.org/wiki/Occam%27s_razor>

Priemyselná revolúcia. [on-line]. [cit. 20. augusta 2020].
Dostupné na internete: <https://sk.wikipedia.org/wiki/Priemyseln%C3%A1_revol%C3%BAcia>

PRISM (NSA) [on-line]. [cit. 23. februára 2022].
Dostupné na internete: <https://sk.wikipedia.org/wiki/PRISM_(NSA)>

Risk management [on-line]. [cit. 3. marca 2022].
Dostupné na internete: <https://en.wikipedia.org/wiki/Risk_management>

Singularita. [on-line]. [cit. 3. augusta 2020].
Dostupné na internete: <https://sk.wikipedia.org/wiki/Singularita>

Social Credit System. [on-line]. [cit. 28. februára 2022].
Dostupné na internete: <https://en.wikipedia.org/wiki/Social_Credit_System>

Stuxnet. [on-line]. [cit. 10. marca 2022].
Dostupné na internete: <https://en.wikipedia.org/wiki/Stuxnet>

Tempora [on-line]. [cit. 23. februára 2022].
Dostupné na internete: <https://sk.wikipedia.org/wiki/Tempora>

The Shadow Brokers. [on-line]. [cit. 10. marca 2022].
Dostupné na internete: <https://en.wikipedia.org/wiki/The_Shadow_Brokers>

Turek (stroj). [on-line]. [cit. 20. augusta 2020].
Dostupné na internete: <https://sk.wikipedia.org/wiki/Turek_(stroj)>

Turing machine. [on-line]. [cit. 7. apríla 2022].
Dostupné na internete: <https://en.wikipedia.org/wiki/Turing_machine>

Turingov test. [on-line]. [cit. 29. januára 2021].
Dostupné na internete: <https://sk.wikipedia.org/wiki/Turingov_test>

Virtual assistant. [on-line]. [cit. 20. augusta 2020].
Dostupné na internete: <https://en.wikipedia.org/wiki/Virtual_assistant>

Whistleblowing [on-line]. [cit. 23. februára 2022].
Dostupné na internete: <https://cs.wikipedia.org/wiki/Whistleblowing>

Wolfgang Kempelen. [on-line]. [cit. 20. augusta 2020].
Dostupné na internete: <https://sk.wikipedia.org/wiki/Wolfgang_Kempelen>

Slovník termínov

Adhocracia	spôsob organizácie spoločenstva ľudí. V *ad hoc* (k určitému účelu vytvorenej) organizácii je každá organizačná zložka modulárna a disponibilná, každá laterálne interaguje s mnohými inými jednotkami. Rozhodnutia sú prijímané podľa okolností, nie štandardným spôsobom.
AGI	skutočná umelá inteligencia, ktorá je všeobecná (general) a silná (strong). Všeobecná, keďže dokáže zvládnuť akúkoľvek intelektuálnu úlohu a má schopnosť generalizovať, t. j. zovšeobecňovať a prenášať, či adaptovať naučené schopnosti na iné úlohy. Silná, pretože aj skutočne rozumie tomu, či rieši a vykonáva.
Algokracia	skrátený názov pre vládu podľa algoritmov, algoritmický právny poriadok, algoritmické vládnutie a pod. Ide o alternatívnu formu vlády, resp. spoločenského usporiadania, pri ktorom sa na reguláciu, presadzovanie práva a všeobecne na akýkoľvek aspekt každodenného života, ako je doprava alebo registrácia pozemkov, používajú počítačové algoritmy, najmä umelá inteligencia a blockchain.
ANI	úzko špecializované systémy umelej inteligencie (narrow AI), ktoré sú optimalizované na zvládnutie konkrétnej úlohy, resp. množiny úloh. Ide súčasne o systémy slabej umelej inteligencie (weak AI), ktoré vykazujú inteligentné správanie na základe modelov a aplikovaných metód i tréningových dát. Hovoríme teda o systémoch, ktoré sú zamerané na riešenie konkrétnych úloh a sú závislé na ľudskom vstupe a konfigurácii.
ASI	Artificial Super Intelligence, t.j. umelá inteligencia, ktorá by bola naprieč všetkými oblasťami inteligentnejšia ako človek. Úvahy o ASI sa mnohokrát spájajú s konceptom singularity v oblasti AI, v ktorej umelá inteligencia so schopnosťou učiť sa, zlepšovať sa a samostatne sa vyvíjať rýchlo dosiahne a následne extrémne prevýši

	inteligenciu človeka. Vo väčšine diskusií sa ASI nespomína samostatne, je vnímaná ako podmnožina AGI.
Cyberspace	virtuálny životný priestor, často nazývaný aj kybernetický priestor, či virtuálny svet, je fiktívny priestor, v ktorom prebieha komunikácia medzi jednotlivými subjektami, pričom tieto sú celkom reálne.
Data mining	(dolovanie dát) je analytický proces navrhnutý na skúmanie veľkého množstva dát podľa konkrétnych vzorov a väzieb. Výsledkom sú množiny dát, splňujúcich žiadané podmienky, prípadne predikcia ďalšieho vývoja, správania, atď.
Digital Age	digitálny vek – obdobie rozvoja spoločnosti po druhej priemyselnej revolúcii, kedy sa na základe využívania IKT spoločnosť mení na spoločnosť založenú na vedomostiach.
Digital divide	digitálne rozdelenie ako dôsledok informačnej nerovnosti, ktorá prerastá do digitálnej chudoby a má evidentné dôsledky na život ľudí.
Digitálna identita	je súbor informácií v elektronickej forme, ktoré reprezentujú subjekt, ktorým môže byť osoba, organizácia alebo objekt reálneho sveta.
eGovernment	elektronická forma výkonu verejnej správy prostredníctvom informačno-komunikačných technológií (IKT). eGovernment je využitie prostriedkov a nástrojov informačných technológií (najmä siete Internet) na skvalitnenie verejných služieb pre občanov, podnikateľov a celú spoločnosť.
eHealth	zdravotnícka starostlivosť podporovaná elektronickými procesmi, informatizáciou a prostriedkami IKT.
Global village	vnímanie spoločnosti ako spoločenského priestoru prirovnávaného celosvetovej dedine. Ide o dôsledok enormného rozvoja využívania prostriedkov IKT, ktoré sú masovo prijímané populáciou, presahujú geografické i kultúrne vzdialenosti a pretvárajú vnímanie spoločnosti, resp. sveta ako jedného spoločného celku.
Haktivizmus	označuje používanie prostriedkov IKT na podporu politickej, či ideologickej agendy. I keď mnohokrát ide o informačnú analógiu protestných akcií, či demonštrácií z reálneho sveta, existujú i početné

	prípady podvratného, či neetického využitia IKT v presadzovaní politickej agendy. Cieľom hacktivizmu často krát býva reálna, či falošná sloboda slova, ľudské práva a sloboda informácií.
Internet	je celosvetový systém prepojených počítačových sietí na základe štandardizovaných pravidiel (súbor protokolov TCP/IP,...). Internet v súčasnosti navzájom prepája milióny rôznych počítačových sietí a v nich prakticky miliardy zariadení, pričom zastrešuje neustále rastúci počet informačných zdrojov a služieb.
Limitovaná AGI	pracovný termín pre všeobecnú a silnú umelú inteligenciu, ktorá síce nie je schopná myslieť ako ľudia, t.j. v celej komplexnosti presahujúcej do roviny vedomia a sebauvedomenia, predsa len myslenie už nesimuluje, ale určitým spôsobom skutočne myslí.
Sieťová neutralita	je princíp, podľa ktorého by poskytovatelia internetových služieb a vlády mali zaobchádzať so všetkými dátami na internete rovnako, nediskriminujúc, či nepreferujúc podľa užívateľa, obsahu, webovej stránky, platformy, aplikácie, typu pripojeného zariadenia, alebo spôsob komunikácie.
Singularita	pojem singularita sa v mnohých oblastiach matematiky a fyziky používa na vyjadrenie stavu, ktorý je zvláštny, nedefinovaný, odchyľujúci sa z trendu, či množiny, stav prekračujúci očakávajúce parametre, nekonečný, či v daných podmienkach neriešiteľný (napr. aktuálne fyzikálne modely, ktorými – veľmi zjednodušene povedané – nedokážeme popísať singulárne body, akými sú napr. Veľký tresk a čierne diery). Pod termínom technologická singularita sa myslí teoretický bod vo vývoji vedeckej civilizácie (znalostnej spoločnosti), v ktorom sa technologický pokrok zrýchli „do nekonečna“ a prevýši všetky predpovede. Singularitou v umelej inteligencii je myslený stav, ktorý nastane, ak umelá inteligencia so schopnosťou učiť sa, zlepšovať sa a samostatne sa vyvíjať rýchlo dosiahne a následne prevýši

	inteligenciu človeka. Teda situácia, keď sa počítačové systémy stanú inteligentnejšími než ľudia.
Singularita silná	pracovný termín pre singularitu v umelej inteligencii, keď super inteligencia (AGI/ASI) prevýši náš intelekt.
Singularita slabá	pracovný termín pre singularitu v umelej inteligencii (ANI), ktorá ovládne a prekoná naše slabosti.
Šifrovanie	je proces, ktorým sa nezabezpečené elektronické dáta prevádzajú pomocou kryptografie na dáta šifrované, čitateľná len tým, kto pozná dešifrovací kľúč, či pravidlo. Šifrovanie dát slúži k ich ochrane pred cudzími osobami a používa sa pri ich ukladaní i prenose prostredníctvom telekomunikačných a informačných kanálov.
World Wide Web	(web 1.0) je distribuovaný informačný systém navzájom prepojených hypertextových dokumentov, ktoré sú dostupné prostredníctvom internetu. Distribuovaný znamená, že jednotlivé dokumenty sa môžu nachádzať kdekoľvek na internete. Hypertextové dokumenty sú dokumenty, ktoré obsahujú časti textu s odkazmi (referenciami) na iné dokumenty, umožňujúc tak priamy prístup k odkazovaným informáciám. Web 2.0 je názov pre súčasné webové technológie charakterizované používateľským vytváraním obsahu a sociálnymi sieťami, širokou technologickou dostupnosťou a webovými aplikáciami, dynamickým i mediálnym obsahom a využívaním v rámci fungovania informačnej spoločnosti. Web 3.0 by mal byť virtuálnym svetom informácií s vysoko interaktívnym a personalizovaným využitím, ktorý bude okrem iných sofistikovaných technológií (napr. tzv sémantický web, virtuálnu realitu a pod.) v plnej miere využívať aj algoritmy umelej inteligencie.

Summary

Digital disruption, emerging information and knowledge society, paradigm shift, permanent technological revolution, etc. are terms exceedingly prominent in the current historical era. In this context, the core of almost every technological innovation takes pride in having artificial intelligence implemented in it. It seems, however, that artificial intelligence is not only a buzzword used to help technology companies to achieve breakthrough and earn money, nor to improve the strategies of the world's leaders. It has become a real concept and will play an integral part in the future of our civilization.

AI systems cause the levels of knowledge and research in various fields to progress rapidly. Moreover, they can be helpful and very useful in almost the whole spectrum of human activity, equally, they are becoming of bigger importance to the life of society.

Nevertheless, the application of AI technology has its dark side and risks. Their failure, misuse or direct exploitation are becoming a nightmare of security, democracy, sociology, and psychology. In order to attain conscious AI systems, the society endeavours to go beyond the current AI systems. This brings about a dilemma, whether this effort will lead to a "golden age" of the civilization's development, or it will become a downfall of all that has been built and achieved for the good of humanity so far.

In this regard, we take the side of voices calling for ethical evaluation, consideration of moral aspects, as well as for defining rules for development, use, and operation of AI systems as such. Due to the progress of information society and the outset of digital age, the questions *if* to use the components of AI technology and *when*, are not taken into consideration because they eventually already exist among us and their presence is expanding constantly. Rather the following question needs to be asked: *how* to use them, i.e. under what conditions, for what purposes, in what manner, and with what consequences the phenomenon of AI should become a part of our world.

With over twenty years of experience in the field of cyber security, we come to realize that AI systems have caused many reasons to worry about implementing new technology and a serious concern for ethical queries like nothing before. The use of #AIEthics as a hashtag and a key word among AI experts is becoming a part of the main flow of AI systems development, implementation, and use within the real world. It is not of peripheral importance for those who are truly experienced in problems associated with AI.

The aim of this publication is to propose the basic principles to be upheld during ethics proposal, execution and usage of any AI systems, based on the analysis of the current state of AI systems development and existing ethical rules.

The study is primarily focused on ethical challenges and moral aspects of modern systems of artificial intelligence referred to under the common term Artificial Narrow Intelligence (ANI). This term covers narrow AI systems which are optimized to perform a specific task, or rather a set of tasks. At the same time, they are called weak AI systems and display intelligent behaviour based on some models, applied methods and training data. The goal of such systems is therefore a solution of specific tasks; they are dependent on interference of humans as well as on human configuration.

A broad and suitably deep interdisciplinary framework proposed is one of a significant contribution of this paper. Without it, no real and successful solution to ethical issues and AI technology challenges would be possible. This framework covers a satisfactory amount of information about technology of the systems as well as the psychological, sociological, and legal aspects of their use.

After having been informed of the nature of this framework, another contribution of the paper is formed: identification, naming, analysis and comprehension of risks connected to AI technologies in respect to using them to the extent possible. It is key to have a broader spectrum of knowledge and complex comprehension of the limits and risks of modern AI systems in order to truly understand the ethical issues of such technologies.

The following distinctive contribution made, is a summary of ethical observations, formed during the analysis of limits and risks of present AI systems.

In an effort to fulfil the purpose of this text we gradually broadened the interdisciplinary framework by analysing the present activities in the AI ethics field, both in examining existing activities and upcoming regulations which aim to secure the ethical ambit of using such technologies.

The main finding of our work is the proposal of the basic ethical principles and guidelines for development, implementing and usage of AI systems. Considering the analysis emphasizing the progressive management of a state, intelligence, global surveillance and AI systems integrated in military, the paper presents our own conclusions, proposals for regulations and recommendations for ethical principles; even for such specific and

important areas listed above.

Several proposals for making the most of the Church's potential are presented, as well as the involvement facilitating the support of ethical approach to the issues with the use of AI in a complete extent of range. Mostly it applies to the mission to unite, guide and promote ethical activities in the world; as well as to a constant effort to emphasize and build universal fraternity and social friendship, even in the digital world and its technologies.

The secondary aspect of our effort is to point to the problems of conscious Artificial General Intelligence (AGI) in respect of current ANI technologies. According to protagonists, AGI should be achievable by strong and general artificial intelligence. The term general stands for its ability to generalize and to cope with any intellectual task; to transfer abilities across tasks or to adapt them for another tasks. The term strong stands for its real understanding of everything it deals and is tasked with.

The fifth chapter discusses various fundamental problems exclusively, though the topics analysed in the previous four chapters provide at least a quick glance beyond the horizon towards strong and general artificial intelligence. Another important contribution of this study is represented by our grasp of the ethical issues of conscious AI (e.g. theory of mind, barrier of meaning, cracks in intelligence, embodiment hypothesis, consciousness, the term "person", implementation of mechanism of conscience, etc.).

Altogether, we present the publication as a novelty in a sense of striving for innovative approach to exploring AI issues with respect to moral theology. Our contribution into the global ethical discourse is comprised of the interdisciplinary understanding of AI phenomenon, the thorough analysis of limits and risks and its ethical conclusion, the proposal of basic ethical principles (general and specific), the reference to arguable factors of conscious AI and the reasons given to support our clear view on comparing it to a human being. This debate is becoming more significant with the evolutionary progress of AI technologies, along with the increasingly growing sophisticated implementation of them into the real world which affects the lives of millions.